Predictive modeling in P&C Insurance

Himchan Jeong, University of Connecticut

1 Introduction

1.1 What is Actuarial Science?

According to wikipedia, Actuarial science is "the discipline that applies mathematical and statistical methods to assess risk in insurance, finance and other industries and professions. 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: Since it has its own traditional education curriculum, which is provided by the Society of Actuaries (SOA) or Casualty Actuarial Society (CAS), the companies have their own developed pricing methods.
- Internal/External Communication: Both senior managers and regulators are reluctant to use of unproven new method which might look like a 'black-box'.
- 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.

1.3 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

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

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.

There are some possible alternatives which can be used for dealing with the pitfalls. For example, Shi et al. (2015) and Garrido et al. (2016) used the observed frequency as a covariate for the average severity as following; $\mathbb{E}\left[\overline{C}|N\right] = e^{X\beta + N\theta}$.

In case of longitudinal property, Boucher et al. (2008) analyzed various method and concluded that random effects model would be a good way for capturing the longitudinal property in P&C insurance.

Finally, we may suggest to use non-traditional method, such as regression with neural network or calculation of credibility premium per each group classified by decision tree.

2 Analysis

2.1 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 5,677 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. Since the dataset includes information on multi-line insurance, here I used building and contents (BC), inland marine (IM), and new motor vehicle (PN) claim information.

Table 1: 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 %	
NoClaimCreditIM	No IM claim in prior year:	Y=1	42.1 %	
NoClaimCreditPN	No PN claim in prior year:	Y=1	10.96 %	
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
CoverageIM	Log coverage amount of IM claim in mm	0	0.85	46.75
lnDeductIM	Log deductible amount for IM claim	0	5.34	9.21
CoveragePN	Log coverage amount of PN claim in mm	0	0.16	25.67

Table 2: Summary statistics for claim frequency

		Minimum	Mean	Variance	Maximum
FreqBC	number of BC claim in a year	0	0.88	37.31	231
FreqIM	number of IM claim in a year	0	0.06	0.1	6
FreqPN	number of PN claim in a year	0	0.16	0.92	19

In terms of frequency, IM has relatively moderate dispersion of the number of claim per year, whereas BC has very wide range. Usually, dataset used to calibrate two-parts GLM in practice rarely contains a policy which has more than six claims in a year. So we may need a different methodology for modelling such unusual high frequency.

Table 3: Distribution of frequency per claim type

Count	BC	IM	PN
0	3993	5441	5360
1	997	182	155
2	333	40	51
3	136	6	33
4	76	4	19
5	31	2	16
6	19	2	13
7	19	0	7
8	16	0	4
9	5	0	4
>9	52	0	15

Table 4: Summary statistics for claim severity

		Minimum	Mean	Variance	Maximum
log(yAvgBC)	(log) avg size of BC claim in a year	5.17	8.76	1.86	16.37
log(yAvgIM)	(log) avg size of IM claim in a year	4.09	8.45	2.23	13.09
log(yAvgPN)	(log) avg size of PN claim in a year	3.56	7.63	1.22	10.71

2.2 Model Specification

For prediction, I applied two type of model, one is likelihood based estimation which includes traditional Poisson-gamma two parts model, and the other is the use of neural network. For frequency part, first I fitted the model with Poisson, zero-inflated Poisson (ZIP), and negative binomial (NB). Since I found that frequency fit with Poisson was best both for BC and IM claim whereas ZIP outperformed all the other candidates in PN claim, I used Poisson in BC and IM, and ZIP in PN, respectively. For severity part, I used usual gamma, generalized pareto (GP), and generalized beta of the second kind (GB2) distribution, which was also used in Yang et al. (2011). Note that GP and GB2 distribution could be derived under the random effects framework for repeated measurement per each policyholder.

More specifically, suppose gamma/inv-gamma random effect model is given as following.

$$Y_t|U \sim \text{ Gamma}(\psi_t, U\frac{\mu_t}{\psi_t}) \quad \text{and} \quad U \sim \text{Inv-Gamma}(\eta+1, \eta)$$

Then We can derive a multivariate joint distribution of $\mathbf{Y}_T = (Y_1, Y_2, \ldots, Y_T)'$ by integrating out the random effects U.

$$\begin{aligned} f_{\mathbf{Y}_T}(\mathbf{y}_T) &= \int_0^\infty \prod_{t=1}^T f_{Y_t|U}(y_t|u)p(u)du \\ &= \frac{\eta^{\eta+1}\prod_{t=1}^T(\psi_t y_t \mu_t^{-1})^{\psi_t}}{(\eta+\sum_{t=1}^T \psi_t y_t \mu_t^{-1})^{\sum\psi_t+\eta+1}} \times \frac{\Gamma(\sum\psi_t+\eta+1)\prod_{t=1}^T y_t^{-1}}{\prod_{t=1}^T\Gamma(\psi_t)\Gamma(\eta+1)} \end{aligned}$$

Now, using given joint density, we may derive conditonal distribution of Y_{T+1} given $\mathbf{Y}_T$. Here, let us denote $w_T = \eta + \sum_{t=1}^T \psi_t y_t \mu_t^{-1}$, and $\eta_T = \eta + \sum_{t=1}^T \psi_t$.

$$\begin{aligned} f_{Y_{T+1}|\mathbf{Y}_T}(y_{T+1}|\mathbf{y}_T) &= f_{\mathbf{Y}_{T+1}}(\mathbf{y}_{T+1})/f_{\mathbf{Y}_T}(\mathbf{y}_T) \\ &= \frac{w_T^{\eta+1}(\psi_{T+1}y_{T+1}\mu_{T+1}^{-1})^{\psi_{T+1}}}{(w_T+\psi_{T+1}y_{T+1}\mu_{T+1}^{-1})^{\psi_{T+1}+\eta_T+1}} \times \frac{\Gamma(\psi_{T+1}+\eta_T+1)y_{T+1}^{-1}}{\Gamma(\psi_{T+1})\Gamma(\eta_T+1)} \end{aligned}$$

As a result, we can see that $Y_{T+1}|\mathbf{Y}_T \sim GP(\eta_T+1, w_T\mu_{T+1}/\psi_{T+1}, \psi_{T+1})$ and $\mathbb{E}\left[Y_{T+1}|\mathbf{Y}_T\right] = \frac{w_T\mu_{T+1}\psi_{T+1}}{(\eta_T+1-1)\psi_{T+1}} = \frac{w_T}{\eta_T}\mu_{T+1}$

Note that we may use the previous argument for the average severity modelling by denoting

$$Y_t = \overline{C}_t|N_t, \ \psi_t = N_t/\phi, \ \mu_t = \exp(X_t\beta + N_t\theta), \ \eta = k/\phi$$

Therefore, we have two types of premium, a priori premium and a posteriori premium, which is a product of weight factor from previous observation and a priori premium under GP distribution.

$$\begin{aligned} \mathbb{E}\left[\overline{C}_{T+1}|N_{T+1}\right] &= \exp(X_{T+1}\beta + N_{T+1}\theta) \\ \mathbb{E}\left[\overline{C}_{T+1}|\overline{\mathbf{C}}_T, \mathbf{N}_T\right] &= \exp(X_{T+1}\beta + N_{T+1}\theta)\frac{k+\sum_{t=1}^T S_t\mu_t^{-1}}{k+\sum_{t=1}^T N_t} \end{aligned}$$

Likewise, we may derive GB2 distribution from random effects framework. Let us assume that G-Gamma/GI-gamma random effect model is given as following. We denote that $z_t = \frac{\Gamma(\psi_t+1/p)}{\Gamma(\psi_t)}$, and $w = \frac{\Gamma(\eta+1)}{\Gamma(\eta+1-1/p)}$.

$$Y_t|U \sim \text{ G-Gamma}(\psi_t, U\frac{\mu_t}{z_t}, p) \quad \text{and} \quad U \sim \text{GI-Gamma}(\eta+1, w, p)$$

Then we can derive a multivariate joint distribution of $\mathbf{Y}_T = (Y_1, Y_2, \ldots, Y_T)'$ by integrating out the random effects U as well.

$$\begin{aligned} f_{\mathbf{Y}_T}(\mathbf{y}_T) &= \int_0^\infty \prod_{t=1}^T f_{Y_t|U}(y_t|u)p(u)du \\ &= \frac{p^T w^{p(\eta+1)}\prod_{t=1}^T(z_t y_t \mu_t^{-1})^{p\psi_t}}{(w^p+\sum_{t=1}^T(z_t y_t \mu_t^{-1})^p)^{\sum\psi_t+\eta+1}} \times \frac{\Gamma(\sum\psi_t+\eta+1)\prod_{t=1}^T y_t^{-1}}{\prod_{t=1}^T\Gamma(\psi_t)\Gamma(\eta+1)} \end{aligned}$$

Now, using given joint density, we may derive conditonal distribution of Y_{T+1} given $\mathbf{Y}_T$. Here, let us denote $w^*_{T,p} = \sqrt[p]{w^p + \sum_{t=1}^{T}(\psi_t y_t \mu_t^{-1})^p}$, and $\eta_T = \eta + \sum_{t=1}^{T}\psi_t$, then we can get

$$\begin{aligned} f_{Y_{T+1}|\mathbf{Y}_T}(y_{T+1}|\mathbf{y}_T) &= f_{\mathbf{Y}_{T+1}}(\mathbf{y}_{T+1})/f_{\mathbf{Y}_T}(\mathbf{y}_T) \\ &= \frac{(w^*_{T,p})^{p(\eta+1)}(z_{T+1}y_{T+1}\mu_{T+1}^{-1})^{p\psi_{T+1}}}{((w^*_{T,p})^p + (z_{T+1}y_{T+1}\mu_{T+1}^{-1})^p)^{\psi_{T+1}+\eta_T+1}} \times \frac{\Gamma(\psi_{T+1}+\eta_T+1)y_{T+1}^{-1}}{\Gamma(\psi_{T+1})\Gamma(\eta_T+1)}. \end{aligned}$$

As a result, we can see that $Y_{T+1}|\mathbf{Y}_T \sim GB2(\eta_T+1, w^*_{T,p}\mu_{T+1}/z_{T+1}, \psi_{T+1}, p)$ so that $\mathbb{E}[Y_{T+1}|\mathbf{Y}_T] = w^*_{T,p}\mu_{T+1}\frac{\Gamma(\eta_T+1-1/p)z_{T+1}}{\Gamma(\eta_T+1)z_{T+1}} = w^*_{T,p}\frac{\Gamma(\eta_T+1-1/p)}{\Gamma(\eta_T+1)}\mu_{T+1}$

Again, we may use the previous argument for the average severity modelling by denoting

$$Y_t = \overline{C}_t|N_t,\ \psi_t = N_t/\phi,\ \mu_t = \exp(X_t\beta + N_t\theta),\ \eta = k/\phi$$

Therefore, we have two types of premium, a priori premium and a posteriori premium as well.

$$\mathbb{E}\left[\overline{C}_{T+1}|N_{T+1}\right] = \exp(X_{T+1}\beta + N_{T+1}\theta)$$

$$\mathbb{E}\left[\overline{C}_{T+1}|\overline{\mathbf{C}}_T, \mathbf{N}_T\right] = \exp(X_{T+1}\beta + N_{T+1}\theta) \times \sqrt[p]{w^p + \sum_{t=1}^{T}(\overline{C}_t\mu_t^{-1}z_t)^p}\frac{\Gamma(k/\phi+1+\sum_{t=1}^{T}N_t/\phi-1/p)}{\Gamma(k/\phi+1+\sum_{t=1}^{T}N_t/\phi)}$$

For a posteriori premium, We may observe that as $k \to \infty$, the following holds.

$$\frac{k+\sum_{t=1}^{T}S_t\mu_t^{-1}}{k+\sum_{t=1}^{T}N_t} \to 1, \qquad \sqrt[p]{w^p + \sum_{t=1}^{T}(\overline{C}_t\mu_t^{-1}z_t)^p}\frac{\Gamma(k/\phi+1+\sum_{t=1}^{T}N_t/\phi-1/p)}{\Gamma(k/\phi+1+\sum_{t=1}^{T}N_t/\phi)} \to 1$$

$$\left(\because \lim_{n\to\infty}\frac{\Gamma(n+\alpha)}{\Gamma(n)n^\alpha} = 1, \quad \alpha \in \mathbb{C} \text{ and } w = \frac{\Gamma(k/\phi+1)}{\Gamma(k/\phi+1-1/p)}\right)$$

Therefore, k works as a smoothing factor for a posteriori premium. In other words, if we choose very small k, then we use more information from the past, whereas if we choose relatively large k, then we use less information from the past. I chose optimal k for each claim using cross-validation in training set and the following is the distribution of weight factors for each claim with optimal k.

Table 5: Distribution of weight factors for each claim

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
BC: GP weight	0.564	0.994	1	1.006	1.000	3.284
BC: GB2 weight	0.394	0.998	1	1.004	1.006	3.756
IM: GP weight	0.718	1.000	1	1.001	1.000	2.268
IM: GB2 weight	0.604	1.000	1	0.998	1.000	1.594
PN: GP weight	0.876	1.000	1	1.000	1.000	1.248
PN: GB2 weight	0.856	1.000	1	1.000	1.000	1.114

Furthermore, I used neural network for both two-parts estimation and direct compound loss estimation. Note that I excluded tweedie distribution since it showed relatively poor performance in terms of validation measures. The following is brief summary for the entertained models. Note that I did not include Tweedie model as a candidate because it underperformed all the other listed models according to a preliminary analysis.

- Likelihood based models
  - Frequency part: Poisson in BC and IM, Zero-inflated
  - Severity part: Gamma, GP (Prior/Posterior), and GB2 (Prior/Posterior)
- Neural network for two-parts and compound loss
- Naive method: sample mean of compound loss in training set

2.3 Estimation and Prediction

For validation measure, I used Gini index and mean squared error (MSE). Gini index, originated from economical concept describing the distribution of wealth among people, is used to distribution of 'risk' among the insured. For detailed explanation, see Frees et al. (2014). I drew the Lorenz curve with the following three-step process computed the corresponding Gini index.

1. From our hold-out sample, for $i = 1, 2, \ldots, M$, sort the observed loss Y_i according to the risk score S_i for which in our case, the predicted value from each model, in an ascending manner. That is, calculate the rank R_i of S_i between 1 and M with $R_1 = \text{argmin}(S_i)$.
2. Compute $F_{\text{score}}(m/M) = \frac{1}{M}\sum_{i=1}^{M} 1_{(R_i \le m)}$, the cumulative percentage of exposures, and $F_{\text{loss}(m/M)} = \frac{\sum_{i=1}^{M} Y_i 1_{(R_i \le m)}}{\sum_{i=1}^{M} Y_i}$, the cumulative percentage of loss, for each $m = 1, 2, \ldots, M$.
3. Plot $F_{\text{score}}(m/M)$ on the x-axis, and $F_{\text{loss}}(m/M)$ on the y-axis.

As we can see, Lorenz curve and Gini index depends only on the 'rank' so if there are many observed loss which have the same value, then it might be distorted. Therefore, I used only positive amounts of actual loss (and corresponding predicted pure premium) for drawing Lorenz curves to avoid that issue.

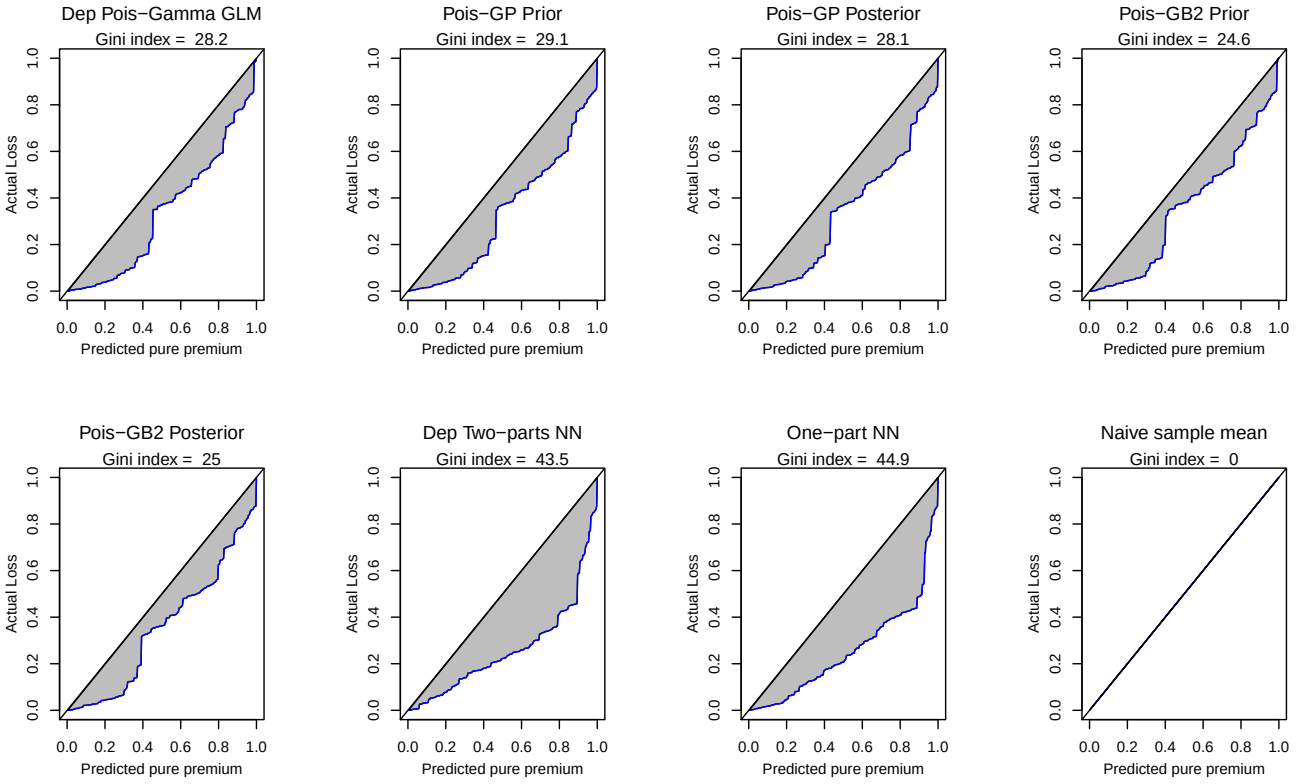

Figure 1: The Lorenz curve and the Gini index values for BC claim

We can see that Gini index for naive model is 0, since we charge the same premium for all policyholder so that there is no risk classification. And one-part neural network showed the best performance in terms of Gini indices in BC and IM claim, while the use of Poisson-GP prior two parts model show the same performance in case of IM claim. Note that for PN claim, there are little differences among the PN models in Gini indices, which might be due to the lack of relevant covariates. Usually, driver's gender, age, vehicle type and capacity have significant effects on the automobile insurance claim and severity. However, in the dataset, there are only regional information which cannot be enough for risk classification of automobile insurance.

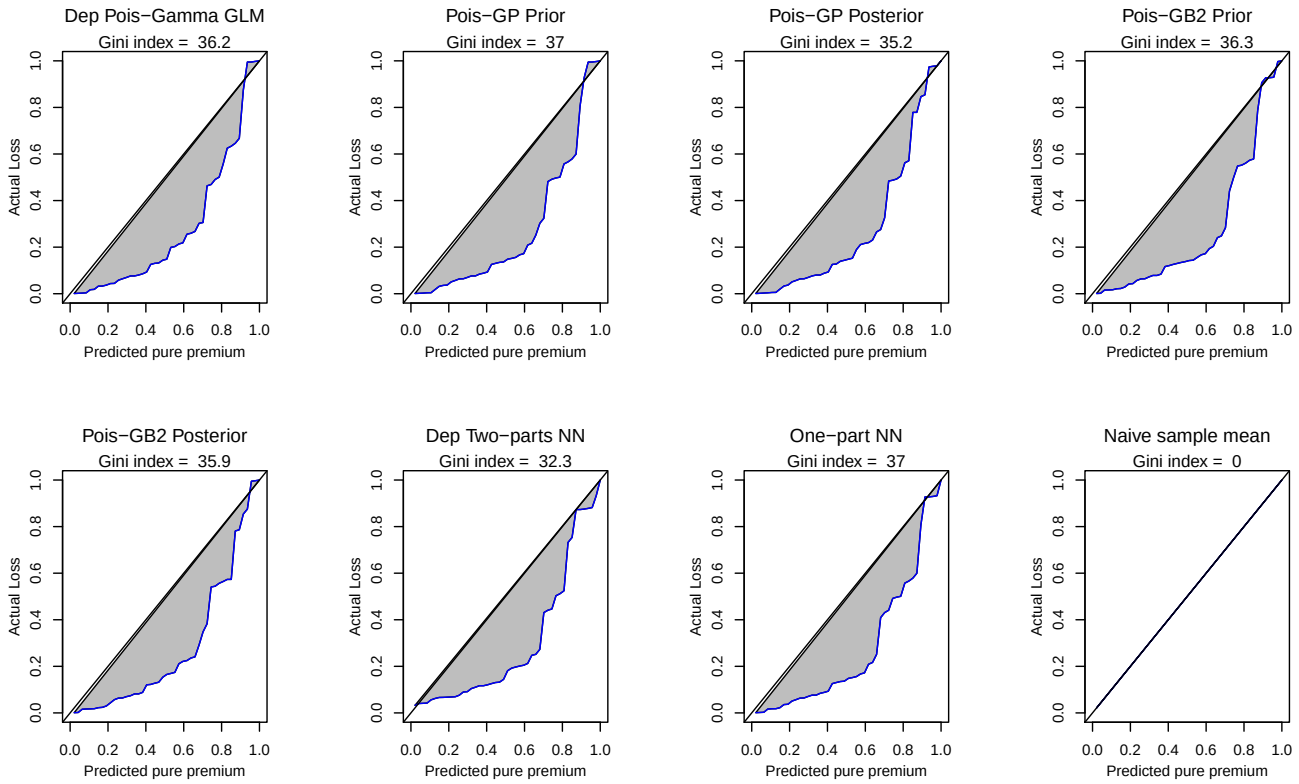

Figure 2: The Lorenz curve and the Gini index values for IM claim

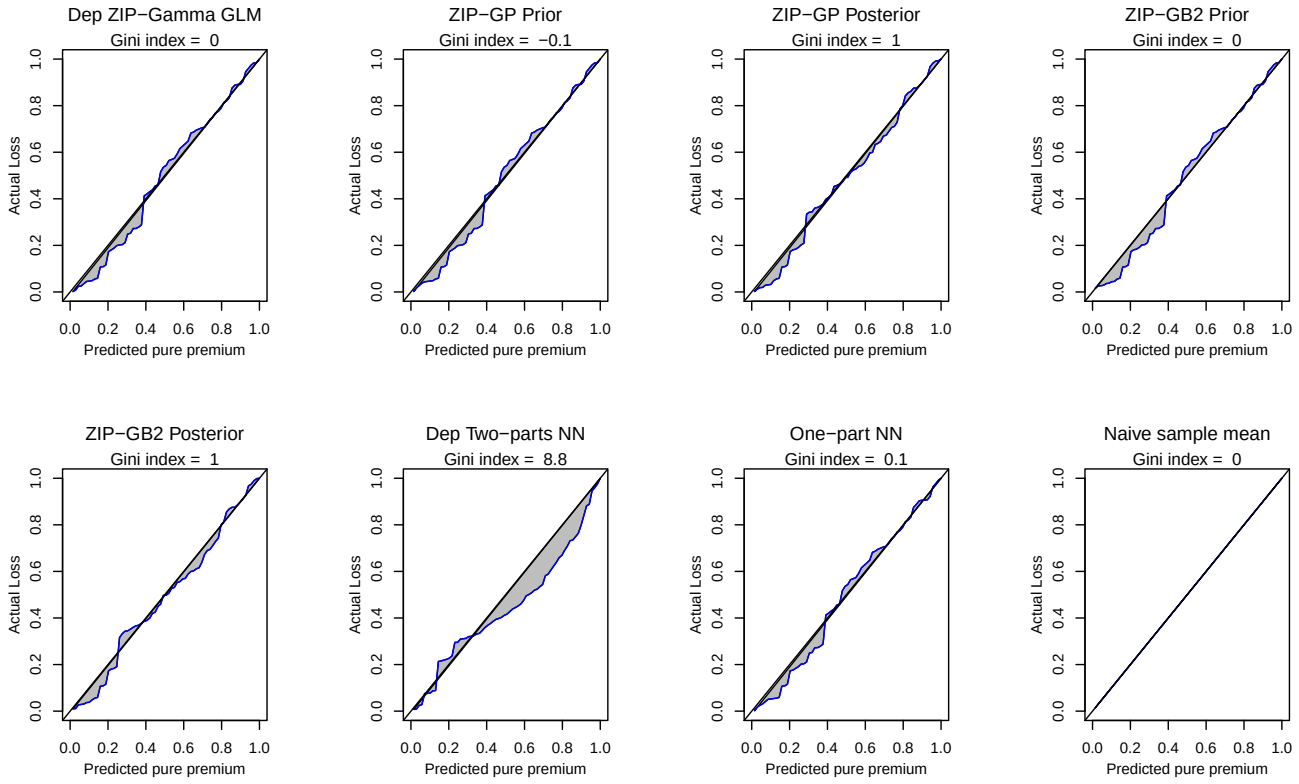

Figure 3: The Lorenz curve and the Gini index values for PN claim

Table 6: MSEs for all type of claim per each model

	BC	IM	PN
Gamma	183466	6647	3673
2P NN	142480	6721	4070
Prior GP	143061	6479	3713
Posterior GP	138416	6445	3716
Prior GB2	139431	6589	3660
Posterior GB2	129824	6510	3662
1P NN	141360	6685	3984
Naive	141367	6696	4051

Sample mean is the most naive and simple estimator but sometimes it is hard to outperform that even with so-called 'sophisticated method' - and that is why insurance companies try to increase market share continually. According to the MSEs of given models, in BC and IM claims there were no big differences among the MSE of naive method, two-parts neural network, and one-part neural network. However, we can see that either posterior GP or posterior GB2 models outperformed naive method and neural network in all types of claim with moderate risk classification level - measured by Gini index. Therefore, I can claim that use of posteriori premium based on GP or GB2 distribution is the best among the entertained models.

3 Concluding Remarks

I could show that with the presence of relevent covariates, use of posterior GB2 distribution showed good performance for the building and contents (BC) claim prediction even with unusual claim feature - very high claim frequency per year. Moreover, in the use of MVGB2 distribution, parameter k works as a regularizing parameter so that $k = \infty$ and $p = 1$ is equivalent to current i.i.d. gamma GLM framework for the average severity. Therefore, proposed MVGB2 is a natural extension of current two-parts model entertained in most of P&C insurance company, which can add the more complexity while retaining interpretability of the model.

As future works, it would be worthwhile to calibrate auto insurance claim with the posterior GB2 distribution, upon the existence of relevant explanatory variables. Furthermore, when deriving GB2 distribution the unit of repetead measurement needs not be limited to each policyholder, but might be the classes of policyholder with the same bonus-malus score, or certain risk homogeneous classes obtained by clustering methods.

References

Boucher, J.-P., Denuit, M., and Guillén, M. (2008). Models of insurance claim counts with time dependence based on generalization of poisson and negative binomial distributions. *Variance*, 2(1):135–162.

Frees, E. W. J., Meyers, G., and Cummings, A. D. (2014). Insurance ratemaking and a gini index. *Journal of Risk and Insurance*, 81(2):335–366.

Garrido, J., Genest, C., and Schulz, J. (2016). Generalized linear models for dependent frequency and severity of insurance claims. *Insurance: Mathematics and Economics*, 70:205 – 215.

Shi, P., Feng, X., and Ivantsova, A. (2015). Dependent frequency-severity modeling of insurance claims. *Insurance: Mathematics and Economics*, 64:417 – 428.

Yang, X., Frees, E. W., and Zhang, Z. (2011). A generalized beta copula with applications in modeling multivariate long-tailed data. *Insurance: Mathematics and Economics*, 49(2):265–284.