Predictive models for P&C insurance

Himchan Jeong

University of Connecticut

Data Science in Action

19 April, 2018

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

Current Approches for Claim Modeling

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

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.

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.

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

Summary Statistics for 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.

Summary Statistics for Frequency (Cont'D)

Table 1: 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

Summary Statistics for 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

MSEs for all Type of Claim per Model (Interim)

MSE	BC	IM	PN
Indep Pois-Gamma	314562.2	12541.825	4349.914
Dep Pois-Gamma	183466.2	6647.310	4349.914
Indep ZIP-Gamma	305939.2	7983.752	3655.829
Dep ZIP-Gamma	203612.1	6939.829	3672.773
Indep-2P NN	142480.4	6720.742	4070.260
Dep-2P NN	142480.2	6720.774	4070.372
1P NN	141360.4	6684.799	3983.987
Tweedie GLM	182278.8	30998.432	4401.668

Entertained Models

- Likelihood based models
  - Frequency part: Poisson in BC and IM, ZIP in PN
  - Severity part: Gamma, Generalized Pareto (GP), and GB2
- Neural network for two-parts / compound loss

GP Distribution

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 $\mathbf{Y}_T = (Y_1, Y_2, \dots, Y_T)'$ by integrating out the random effects U.

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

Conditional Distribution from the MVGP

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 $\eta_T = \eta + \sum_{t=1}^{T} \psi_t$.

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

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 $\mathbb{E}\left[Y_{T+1}|\mathbf{Y}_T\right] = \frac{w_T \mu_{T+1} \psi_{T+1}}{(\eta_T+1-1)\psi_{T+1}} = \frac{w_T}{\eta_T}\mu_{T+1}$

A Posteriori Premium for Average Severity in GP

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.

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

GB2 Distribution

G-Gamma/GI-gamma random effect model is given as following. Let us 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 $\mathbf{Y}_T = (Y_1, Y_2, \dots, Y_T)'$ by integrating out the random effects U as well.

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

Conditional Distribution from the MVGB2

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

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

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
$\mathbb{E}\left[Y_{T+1}|\mathbf{Y}_T\right] = w^*_{T,p}\mu_{T+1}\frac{\Gamma(\eta_T+1-1/p)z_{T+1}}{\Gamma(\eta_T+1)z_{T+1}} = w^*_{T,p}\frac{\Gamma(\eta_T+1-1/p)}{\Gamma(\eta_T+1)}\mu_{T+1}$

A Posteriori Premium for Average Severity in GB2

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.

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

Remark for the weight factor in a Posteriori Premium

We may observe that as $k \to \infty$, the following holds.

$$
\begin{aligned}
&\frac{k + \sum_{t=1}^{T} S_t \mu_t^{-1}}{k + \sum_{t=1}^{T} N_t} \;\to\; 1, \\
&\sqrt[p]{w^p + \sum_{t=1}^{T} (\overline{C}_t \mu_t^{-1} z_t)^p} \frac{\Gamma(k/\phi + 1 + \sum_{t=1}^{T} N_t/\phi - 1/p)}{\Gamma(k/\phi + 1 + \sum_{t=1}^{T} N_t/\phi)} \;\to\; 1 \\
&\left(\because \;\lim_{n\to\infty} \frac{\Gamma(n+\alpha)}{\Gamma(n)n^\alpha} = 1, \quad \alpha \in \mathbb{C} \;\text{ and }\; w = \frac{\Gamma(k/\phi + 1)}{\Gamma(k/\phi + 1 - 1/p)}\right)
\end{aligned}
$$

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.

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    1.00 3.28
## BC: GB2 weight 0.394   0.998      1 1.004    1.01 3.76
## IM: GP weight  0.718   1.000      1 1.001    1.00 2.27
## IM: GB2 weight 0.604   1.000      1 0.998    1.00 1.59
## PN: GP weight  0.876   1.000      1 1.000    1.00 1.25
## PN: GB2 weight 0.856   1.000      1 1.000    1.00 1.11
```

Validation Measures for Model Comparison

- Mean Squared Error
- Gini Index
  - Since total claim amounts in validation set are mostly 0, use of gini index could be misleading.
  - So I used only positive amounts of actual loss (and corresponding predicted pure premium) for drawing Lorenz curves.

Gini Indices for BC Claim

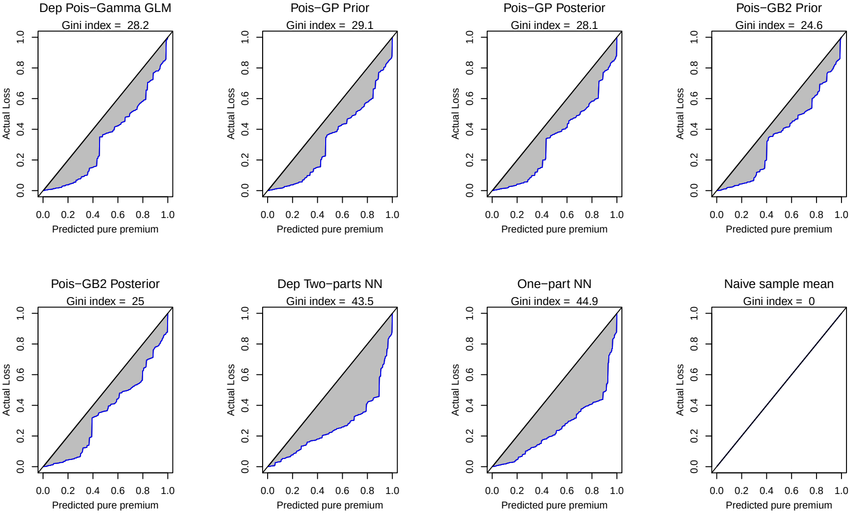

Figure 1: The Lorenz curve and the Gini index values for BC claim

Gini Indices for IM Claim

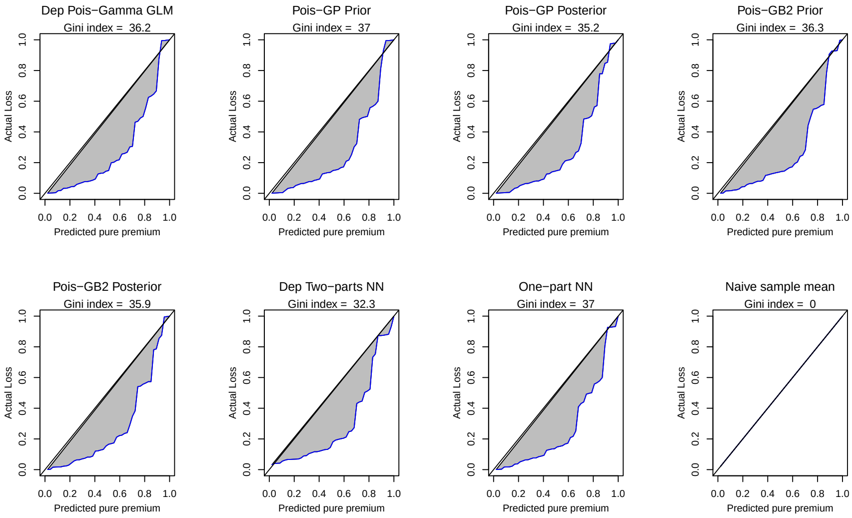

Figure 2: The Lorenz curve and the Gini index values for IM claim

Gini Indices for PN Claim

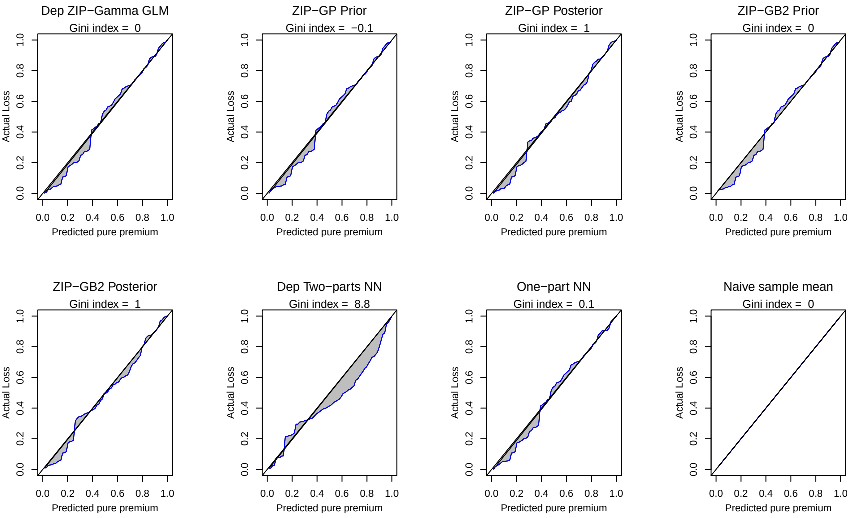

Figure 3: The Lorenz curve and the Gini index values for PN claim

MSEs for all Type of Claim per Model

```
##                         BC       IM       PN
## Gamma            183466.2 6647.310 3672.773
## 2P NN            142480.2 6720.774 4070.372
## Prior GP         143061.0 6478.517 3712.856
## Posterior GP     138416.2 6445.320 3716.482
## Prior GB2        139431.2 6588.580 3659.765
## Posterior GB2    129824.3 6510.384 3661.646
## 1P NN            141360.4 6684.799 3983.987
## Naive            141366.8 6695.585 4051.355
```

Analysis of the Results

- 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 MSE of given models, in all claim use of posteriori premium based on MVGP or MVGB2 distribution outperformed all the other models.
- In case of PN claim model, every model showed poor performance for risk classification, which might be due to the lack of relevant explanatory variables.

Concluding remarks

- 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.
- 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.

Future Works

- It would be worthwhile to calibrate auto insurance claim with the posterior GB2 distribution, upon the existence of relevant explanatory variables.
- For deriving MVGB2 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.