An Extension of Generalized Linear Models for dependent frequency and severity

Masar Al-Mosawi

October 12, 2018

Länsförsäkringar Fondliv

Table of contents

1. Introduction
2. Model building
3. Inference
4. Generalized Linear Models
5. Generalized Linear Models extension
6. Results
7. Conclusion

Introduction

Introduction

In non-life pricing the pure premium is modeled as the product of the two estimates: Claim frequency and claim severity. A general problem is that the frequency and severity are traditionally assumed to be independent.

This assumptions is not always vindicated, car insurance policyholders who tend to file several claims per year are often associated with lesser claim amounts than policyholder who tend to file lesser claims per year.

There is thus a need to account for potential association between claim frequency and claim severity. In this thesis we will construct and analyze the classical model, and a proposed extension of the classical model where claim frequency and claim severity are dependent.

Model building

Model building

Variations can be estimated by a set of covariates. The range for each covariate are called classes. Let M be the number of covariates, and let m_i be the number of classes for covariate i. A tariff cell is denoted by the vector $(i_1, \ldots, i_M)$. We use the multiplicative model for the expected value of a response variable Y:

$$E[Y_{i_1,\ldots,i_M}] = \mu_{i_1,\ldots,i_M} = \gamma_0 \gamma_{1i_1} \gamma_{2i_2} \ldots \gamma_{Mi_M}, \tag{1}$$

where, the γ is called the relativities. The relativities measure the effect when all other variables are held constant

Model building

Generalized Linear Models (GLMs) is a class of statistical methods which generalizes the linear models. GLM solves two problems that occurs with linear models when applying it to non-life insurance pricing:

- GLM assumes general class of distribution instead of normal distribution
- GLM has a link function instead of the mean being a linear function. Multiplicative model is more reasonable for pricing

GLMs uses Exponential Dispersion Models (EDMs) that generalize the normal distribution, that are used in linear models, into a family of distributions for the GLMs.

$$f_{Y_i}(y_i, \theta_i, \phi) = exp\left\{\frac{y_i\theta_i - b(\theta_i)}{\phi/w_i} + c(y_i, \phi, w_i)\right\}, \tag{2}$$

Inference

Inference

To estimate the parameters in GLM we use the maximum-likelihood estimation (ML):

The method of maximum likelihood is based on the log-likelihood function $l(\theta, \phi, y)$, which is a function of the parameters of a statistical model.

- Given a family of distributions, the method of ML finds the values of the model parameter θ, that maximize the log-likelihood function
- Intuitively, the ML selects the parameters that make the data y most probable

Inference

For testing a ML-estimated parameters significance, we use the null hypothesis method:

The null hypothesis method is the use of statistics to determine the probability that a given hypothesis is true

1. Formulate the null hypothesis $\theta = \theta_0$
2. Identify a test statistic that can be used to assess the truth of the null hypothesis
3. Compute the p-value, which is the probability that a test statistic at least as significant as the one observed would be obtained assuming that the null hypothesis is true. The smaller the p-value, the stronger the evidence against the null hypothesis
4. Compare the p-value to an acceptable confidence level $1 - \alpha$. If $p \leq \alpha$, the null hypothesis is rejected

Inference

In GLM a generalization of the idea of using the sum of squares of residuals for a good measure of goodness-of-fit is the deviance function. It can asses which model fits the data best.

$$D(y, \mu) = 2(l(\theta, \phi, y) - l(\theta, \phi, \mu)). \tag{3}$$

- The saturated model is used as a benchmark in measuring the goodness-of-fit of other models, since it has the perfect fit
- One can view the deviance function as a distance between two probability distributions and can be used to perform model comparison
- The deviance functions will generate deviance plots for model validation, they can asses which model fits the data best

Inference

Another criteria for estimating the quality of models in purpose for model selection is the Akaike information criteria (AIC).

$$AIC = -2l(\hat{\theta}, \phi, y) + 2K, \tag{4}$$

- AIC rewards goodness of fit (as assessed by the log-likelihood function), but it also includes a penalty that is an increasing function of the number of estimated parameters
- In other words, AIC value is used to determine which model minimizes the loss of information when approximating reality given the data at hand
- $\Delta_i = AIC_i - AIC_{min}$ is a measure of each model relative to the best model

Generalized Linear Models

Generalized Linear Models

For a fixed time period $w = 1$, the total amount paid out in claims is: $S = \sum_{j=1}^{N} Y_j$. S the total amount paid out in claims, N is the number of claims, Y_j is the claim amount for the jth incurred claim.

Assuming that the claim frequency and claim severity is independent: $E[S] = E[N]E[Y]$.

- The number of claims is assumed to be poisson distributed, $N \sim P(v_i)$
- The claim amount is assumed to be gamma distributed, $Y \sim G(\alpha, \beta)$

The poisson distribution and the gamma distribution are members of the EDM family.

Generalized Linear Models

For number of claims N_i, let $v_i = E[N_i]$. Then:

- The ML-equations: $\sum_i x_{ij}(n_i - v_i) = 0$.
- The deviance function: $D(n, v) = 2\sum_i(n_i log(n_i/v_i) + (v_i - n_i))$.

For claim amount Y_i, let $\mu_i = E[Y_i]$. Then:

- The ML-equations: $\sum_i \frac{x_{ij}}{\mu_i}(y_i - \mu_i) = 0$.
- The deviance function: $D(y, \mu) = 2\sum_i(-1 + \frac{y_i}{\mu_i} + log(\frac{\mu_i}{y_i}))$.

Generalized Linear Models extension

Generalized Linear Models extension

For a fixed time period $w = 1$, the total amount paid out in claims is: $S = \sum_{j=1}^{N} Y_j$. S the total amount paid out in claims, N is the number of claims, Y_j is the claim amount for the jth incurred claim.

To account for dependence, the mean of the severity distribution is allowed to depend on N

$$E[S] = E[NE[\overline{Y}|N]], \tag{5}$$

where $\overline{Y}|N = (Y_1 + \cdots + Y_N)/N$ is the average claim severity, S is the aggregate losses incurred and N is the number of claims.

Generalized Linear Models extension

Two reflections on the dependent setup:

- Claim count N is modeled in exactly the same way as in the classical GLM approach.
- The average claim severity $\overline{Y}$ using claim N as both covariate in the GLM, and weight factor in the EDM.

One has $E[S] = E[NE[\overline{Y}|N]] \neq E[N]E[Y]$.

- Independence: $E[S] = E[N]E[Y] = v\mu$
- Dependence: $E[S] = E[NE[\overline{Y}|N]] = v\mu e^{v(e^{\theta}-1)+\theta}$

An dependence factor emerges: $e^{v(e^{\theta}-1)+\theta}$, together with a dependence parameter θ. It is the estimate of the covariate N.

Generalized Linear Models extension

For number of claims N_i, let $v_i = E[N_i]$. Then:

- The ML-equations are same as in the classical GLM: $\sum_i x_{ij}(n_i - v_i) = 0$.
- The deviance function is same as in the classical GLM: $D(n, v) = 2\sum_i (n_i log(n_i/v_i) + (v_i - n_i))$.

For average claim severity $\overline{Y}_i$, let $\mu_{\theta i} = E[\overline{Y}_i]$. Then:

- The ML-equations: $\sum_i^m \frac{n_i x_{ij}}{\mu_{\theta i}}(\overline{y}_i - \mu_{\theta i}) = 0$.
- Additional ML-equations: $\sum_i^m \frac{n_i^2}{\mu_{\theta i}}(\overline{y}_i - \mu_{\theta i}) = 0$.
- The deviance function: $D(y, \mu) = 2\sum_i^m n_i(-1 + \frac{\overline{y}_i}{\mu_{\theta i}} + log(\frac{\mu_{\theta i}}{\overline{y}_i}))$.

Results

Results

Data from the former Swedish insurance company Wasa, and concerns partial casco insurance for motorcycles.

Covariates	Description	Classes
Zon	Geographic zone	(1,2,3,4,5)
MC class	Mc class	(1,2,3,4)
Vehicle age	The vehicle age	(1,2,3,4)

Table 1:

Claim count	Frequency	Percent	Average amount (Kr)
0	412	67 %	0
1	178	29%	83 372
2	26	4% (13%)	84 674

Table 2:

Results

The dependence parameter θ was estimated to $\hat{\theta} = -0.3472$. The null hypothesis method yields:

1. The null hypothesis $H_0 : \hat{\theta} = \theta_0 = 0$
2. A statistic is identified as the test statistic for the underlying distribution.
3. p-value $= 0.0245$
4. Hence we reject the null hypothesis on confidence level of 97.5% with a $\alpha = 0.0250$, since $p < \alpha$

Results

For the GLM extension, the AIC value is computed to:

- $AIC_{min} = 2637$
- but when we drop the claim count as an covariate the AIC value increases to $AIC_i = 2641$
- $\Delta_i = AIC_i - AIC_{min} = 4$

Results

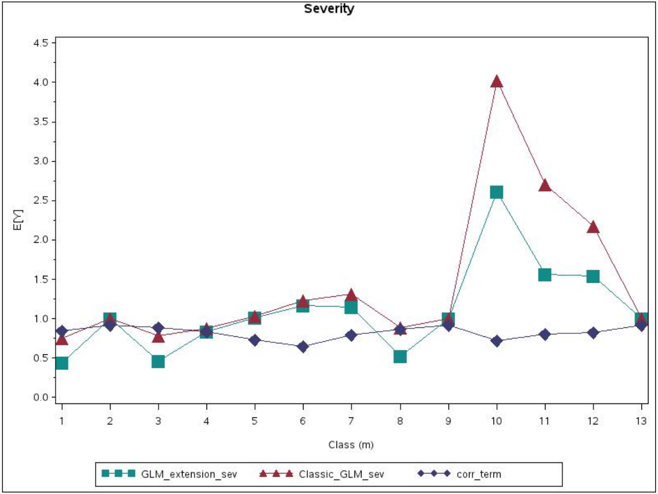

Figure 1: Comparison of the claim severity between the classic GLM and the GLM extension.

Results

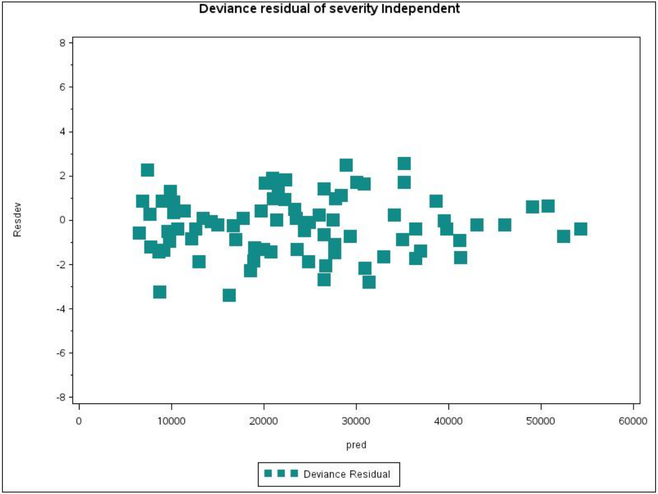

Figure 2: The deviance of the claim severity for the classical GLM.

Results

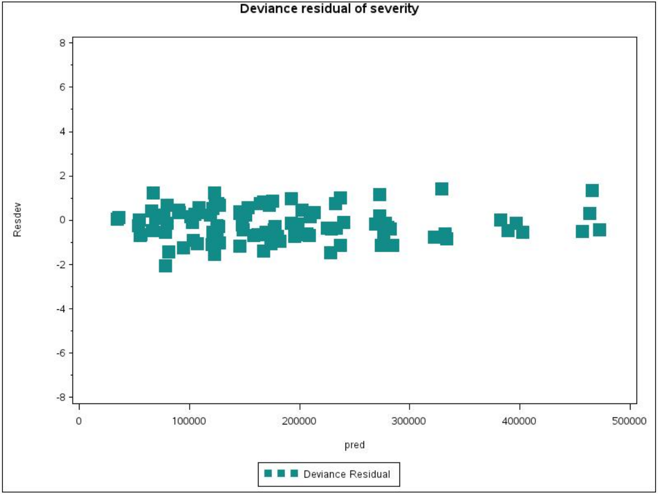

Figure 3: The deviance of the claim severity model for the GLM extension.

Conclusion

Conclusion

- Claim count is a significant covariate for the GLM extension.
- $\Delta_i = 4$ indicates that GLM extension model with claim count is the better model, than without the claim count. But it is not big enough to fully accept claim count as a covariate.
- Deviance figure for the severity has a lower variance, showing that the GLM extension model fit the observations better than the classical GLM.

- Small data to fully confirm that the GLM extension is the better model than the classic GLM, but we have strong evidence to support it.
- The structure for the dependence approaches makes it very easy to implement
- Further studies can be made with greater data and different distributions on claim count and claim amount

Thank you!

masar.al-mosawi@lansforsakringar.se