São Paulo 2025

THE ROLE OF COVARIATES IN CYBER RISK RATEMAKING USING GAMLSS

ALANA AZEVEDO

FEDERAL UNIVERSITY OF CEARÁ / ASSOCIATE PROFESSOR

AIMS AND MOTIVATION

Cyber risks, with other sorts of operational risks, have become a recurrent topic when it comes to proper management.

As this is a challenge in relation to the risk classification and prediction of loss amounts, there is an interest in studying techniques that can satisfactorily manage this risk category.

AIMS AND MOTIVATION

Our main goal is to seek for an actuarial model for the coverage of cyber risks losses using all available information in the estimation of aggregate loss distribution.

We propose an analysis of cyber risk losses, introducing a framework of the GAMLSS which can model not only the mean but all other parameters.

GAMLSS is a general framework for fitting regression type models that include highly skew and kurtotic continuous and discrete distributions.

AIMS AND MOTIVATION

We intend to identify significant risk classification variables, determine tariff classes and calculate premiums considering an a priori model.

We introduce a new perspective to the study of cyber risk pricing with GAMLSS. The particular strengths and differential of the current study, are: (i) Within the framework of the GAMLSS, the use of risk classes in order to compare with the Loss Distribution Approach ratemaking process to check the differences in tariff values, and (ii) The use of real data from a world collection of publicly reported operational losses.

CYBER RISK DATA

Description

CYBER RISK DATA DESCRIPTION

For our empirical analysis of cyber risk we rely on the SAS® OpRisk Global Data which is the world's largest collection of publicly reported operational losses. The database gives information about 37,429 incidents of operational loss in the period between March 1971 and January 2021.

CYBER RISK DATA DESCRIPTION

For each incident, the database reports:

- The description of the event;
- The business lines and industry sectors;
- The risk category;
- Country of incident;
- The amount of the loss and other informations about the firms involved.

All losses, given in US$, are presented in current value for proper comparison.

CYBER RISK DATA DESCRIPTION

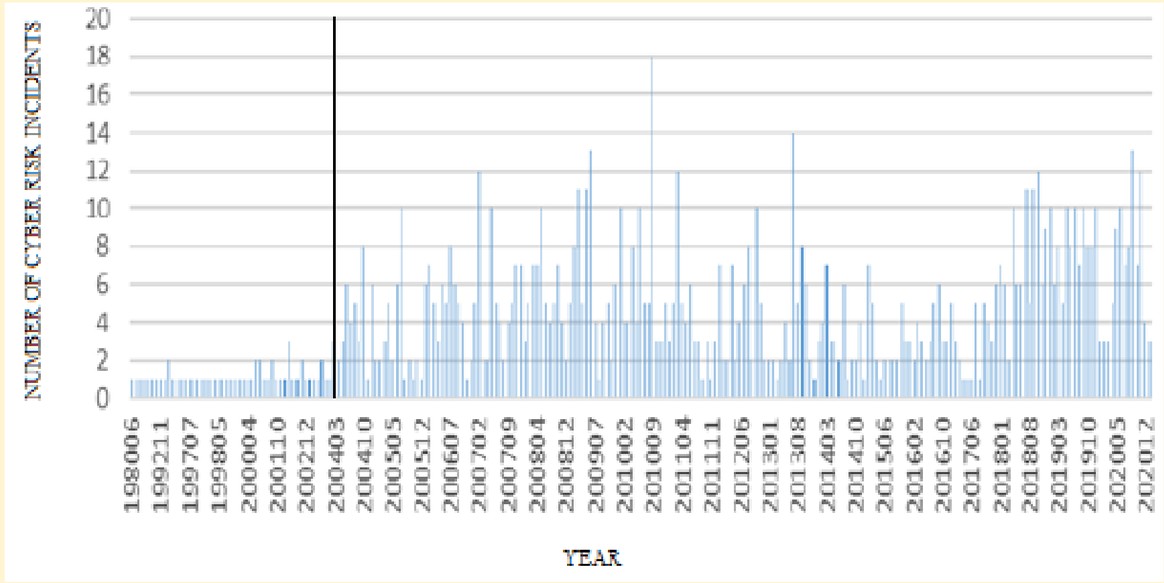
Regulators of insurance and financial markets categorize cyber risk as operational risk, to identify cyber risk in the database, we considered the categorization of CRO (2016) that enumerates the following, we quote:

- Any risks emanating from the use of electronic data and its transmission, including technology tools such as the internet and telecommunications networks;
- Physical damage that can be caused by cyber attacks;
- Fraud committed by misuse of data;
- Any liability arising from data use, storage and transfer;
- The availability, integrity and confidentiality of electronic information (be it related to individuals, companies or governments).

CYBER RISK DATA DESCRIPTION

We decided to consider two subcategories for cyber risk: (1) Actions of people, and (2) Systems and technical failure. Considering information of the SAS® database with complete records from 2004, a total of 974 cyber risk incidents were identified.

Figure 1 – Cyber risk incidents



CYBER RISK DATA DESCRIPTION

Some statistics:

- In 89.6% of the cases, human behavior is the main source of cyber risk incidents;
- Losses by systems and technical failure are, in million US$, 21.71 greater than the total average loss amount;
- North America presents more than half of the incidents (53.6%);
- Finantial service industries hold 70.1% of the total cyber risk incidents.

THE LOSS DISTRIBUTION APPROACH

Ratemaking

THE LOST DISTRIBUTION APPROACH

Karam (2014) explains that LDA consists of separately estimating a frequency distribution for the occurrence of losses and a severity distribution for the economic impact of the losses. After established these two distributions we combine both to obtain an aggregate loss distribution.

For modelling the claim frequency, we denote N as the number of claims over the time period and X_i the i-th claim severity. To model individual claim severity, we denote by X, indexed X_i is the i-th severity, it is necessary to assume that all losses are considered positive, independent and identically distributed random variables from a continuous distribution.

THE LOST DISTRIBUTION APPROACH

Table 1 – Claim frequency goodness-of-fit

Distribution	df	AIC
Generalised Poisson	2	974.26
Negative Binomial	2	974.90
Double Poisson	2	986.79
Zero Adjusted Poisson	2	1,025.76
Zero Inflated Poisson	2	1,028.53

Table 2 – Claim severity goodness-of-fit

Distribution	df	AIC
Weibull	2	2,146.50
Generalised Pareto	2	2,153.60
Log Normal	2	2,154.04
Gamma	2	2,161.42
Inverse Gaussian	2	2,303.41

We considered the Akaike Information Criterion (AIC), a method that allows comparing models with different families of distributions and that does not need further inferences about the model to corroborate its result, we refer to (Burnham and Anderson, 2004).

THE LOST DISTRIBUTION APPROACH

By the classic actuarial method of premium calculation, the net premium is given by the expected value of the insurer's payout per time unit E(S), where the random variable S represents the aggregate amount of claims arising from cyber risk, or generated by the portfolio for the period under study.

$$\begin{aligned} S &= X_1 + X_2 + \ldots + X_N \\ E(S) &= E(N)E(X)\,. \end{aligned}$$

Table 3 – Result of the fitted GP Distribution

	Estimate	Std.Error	t value	Pr($>\|t\|$)
α	1.6064	0.0444	36.1712	< 2.22e-16 ***
β	-2.5347	0.1870	-13.5521	< 2.22e-16 ***

Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 · 0.1 · 1

Table 4 – Result of the fitted Weibull Distribution

	Estimate	Std.Error	t value	Pr($>\|t\|$)
λ	4.2683	0.1217	35.0776	< 2.22e-16 ***
δ	-0.4829	0.0537	-8.9934	< 2.22e-16 ***

Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 · 0.1 · 1

$$\begin{aligned} E(N) &= \frac{1.6064}{(1+2.5347)} = 0.4545 \\ E(X) &= 4.2683\,\Gamma\left(1 - 10.4829^{-1}\right) = 58.9004 \\ E(S) &= (0.4545)(58.9004) = 26.7702 \end{aligned}$$

THE LOST DISTRIBUTION APPROACH

We fit a compound GP distribution with a Weibull as secondary distribution, estimated parameters are in Tables 3 and 4, so that corresponding means come, with estimated net premium as 26.77 (in million US$).

GAMLSS

Ratemaking

GAMLSS

The Generalised Additive Models for Location, Scale and Shape (briefly GAMLSS) is a general framework for fitting regression type models that include highly skew and kurtotic continuous and discrete distributions.

It is a framework that consider a single response variable, allowing many explanatory valuables. The dependence of the response variable in relation to the explanatory variables could be linear, non-linear parametric function or non-parametric smoothing functions.

GAMLSS

Another feature that differs this model from linear (LM), generalized linear (GLM) and generalized additive (GAM) models is that the assumed distribution of the response variable can belong to any parametric distribution, not just to the exponential dispersion family.

GAMLSS

Considering information of the SAS® database with complete records, i.e. with availability of all the explanatory variables under consideration, we developed a GAMLSS model with an application to insurance ratemaking.

There were 680 policies that met our criteria. This subsection describes the modelling results of the best fitted distributions/GAMLSS models that have been applied to model claim frequency and severity.

GAMLSS

The *a priori* rating variables we employ are:

- The type of industry (T), T1 for Financial services and T2 for Non-financial);
- The size of the company (S), S1 as Small, S2 as Medium and S3 as Large;
- The geographic region (R), R1 for Asia, R2 for Europe, R3 for North America and R4 for Other.

GAMLSS

Table 5 presents the best fitted distribution/GAMLSS model for approximating the number of claims. The Zero Modified Logarithmic distribution (ZALG) GAMLSS model was chosen considering the ML estimators of the parameters associated with the condition of significance for all combinations of the covariates.

Table 5 – Result of the fitted ZALG GAMLSS model

Variable	Coeff. β	Std.Error	t value	Pr($>\|t\|$)
Intercept	-0.7087	0.3555	-1.994	0.046582*
T				
T_1	0.7972	0.1623	4.913	1.13e-06***
S				
$S_1 + S_2$	-0.9349	0.1567	-5.968	3.88e-09***
R				
R_3	1.0340	0.2973	3.478	0.000538***
$R_2 + R_4$	0.4879	0.2854	1.709	0.087832 .
AIC	1,078.05			

Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1

GAMLSS

Table 6 presents the best fitted distribution/GAMLSS model for approximating the severity of claims. The Gamma (GA) GAMLSS model was chosen considering the ML estimators of the parameters associated with the condition of significance for all combinations of the covariates.

Table 6 – Result of the fitted GA GAMLSS model

Variable	Coeff. β	Std.Error	t value	Pr($>\|t\|$)
Intercept	4.6303	0.2508	18.465	< 2e-16***
T				
T_1	-0.7906	0.1505	-5.253	2.01e-07***
S				
$S_1 + S_2$	-0.4399	0.1698	-2.590	0.00980**
R				
R_3	-0.4934	0.2024	-2.438	0.01502*
$R_2 + R_4$	-0.5929	0.2177	-2.724	0.00663**
AIC	4,764.63			
Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1				

GAMLSS

Table 7 contains a detailed description of the estimated coefficients for the adjusted GAMLSS, including the tariff relativities associated with each of the risk levels of the variables considered, relativities that express in which direction and in what intensity the statistical premium should be increased or smoothed.

Table 7 – Coefficients, β, and relativities, exp(β), for estimated GAMLSS

		Frequency (ZALG)		Severity (GA)		A priori premium	
Risk factor	Level	β_F	$\exp(\beta_F)$	β_S	$\exp(\beta_S)$	$\beta_F+\beta_S$	$\exp(\beta_F)\exp(\beta_S)$
Intercept	-	-0.7087	0.4923	4.6303	102.5448	3.9216	50.4812
Industry	2	0.0000	1.0000	0.0000	1.0000	0.0000	1.0000
	1	0.7972	2.2193	-0.7906	0.4536	0.0066	1.0066
Size	3	0.0000	1.0000	0.0000	1.0000	0.0000	1.0000
	1+2	-0.9349	0.3926	-0.4399	0.6441	-1.3748	0.2529
Region	1	0.0000	1.0000	0.0000	1.0000	0.0000	1.0000
	3	1.0340	2.8123	-0.4934	0.6105	0.5406	1.7170
	2+4	0.4879	1.6289	-0.5929	0.5527	-0.1050	0.9003

GAMLSS

The relativities were obtained using the inverse exponential function, considering that the link function used in the adjustment of the GAMLSS was logarithmic.

This measure aims to indicate the chance or the marginal effect of the risk observed in relation to the dependent variable when there are variations or changes in the behavior of the realizations of one of the independent variables.

GAMLSS

Regarding the estimated frequency relativities:

The expected average number of claims per policy is higher for companies at tariff Level 1 of the industry type variable, when compared to companies at Level 2 of the same variable.

Thus, it is estimated that the average number of claims to be observed for Level 1 is approximately 2.2193 times the number observed for Level 2, or that the average number of claims observed for Level 1 is, approximately 121.93% higher than the number observed for Level 2.

GAMLSS

Regarding the estimated severity relativities:

The relativity of risk Level 1 and 2 policies in relation to base risk Level 3 policies, in the size variable, is 0.6441.

Thus, it is estimated that the average severity of claims to be observed for Levels 1 and 2 is approximately 0.6441 times the severity observed for Level 3, or that the severity of claims observed for Levels 1 and 2 is approximately 35.59% lower than the average severity observed for Level 3.

GAMLSS

Regarding the estimated *a priori* premium relativities:

The relativity of risk Levels 2 and 4, in relation to that of base risk Level 1, of the region variable, is 0.9003.

This implies that the premium to be paid by Levels 2 and 4 policyholders must be equal to 0.9003 times the amount paid by Level 1 policyholders, or that the premium paid by Levels 2 and 4 policyholders will be reduced by approximately 9.97% in relation to the premium paid by Level 1 policyholders.

GAMLSS

Based on the use of the net premium calculation principle, we analysed the premium for each of the 12 different risk classes and their levels, which are determined by the relevant a priori characteristics.

To calculate the premium of any insured, given their individual risk profile, with N ~ ZALG(μ; σ) and X ~ GA(α; β), without loss of generality one has that the pure premium for a given policy i (Risk S_i and premium P_{Ri}) can be calculated as

$$P_{R_i} = E[S_i] = E[N_i]E[X_i]$$

$$ln(P_{R_i}) = ln(E[S_i]) = \beta_0 + \beta_{11}X_{11} + \beta_{12}X_{12} + \beta_{21}X_{21} + \cdots + \beta_{33}X_{33},$$

GAMLSS

Considering for example an insured with a risk profile categorized by being a financial company, small or medium and located in North America with estimates taken from Table 7, seventh column, the pure premium would total (in million US$)

$$P_{R_i} = e^{3.9216+0.0066-1.3748+0.5406} = 22.07.$$

The group with the lowest expected rate are those small and medium companies in nonfinancial services located in Europe and Other, with a rate of 11.49 (in million US$). On the other hand, the group with the highest expected rate are those large companies in financial services located in North America, with a rate of 87.25 (in million US$).

REMARKS

Conclusions

REMARKS AND CONCLUSIONS

Our intention in this work was to carry out a comparison between the results generated by the Loss Distribution Approach (LDA) and by the Generalized Additive Models of Location, Scale and Shape (GAMLSS) in the treatment of cyber risk data.

In both approaches, frequency and severity of claims were treated separately.

REMARKS AND CONCLUSIONS

For the LDA, covariates were not considered, thus, the individual characteristics of each company were disregarded, generating a single tariff for the portfolio, 26.77 (in million US$).

The GAMLSS model generated results for 12 risk classes resulting from the combination of the considered covariates, these: type of industry (two possibilities), geographic region (four possibilities) and company size (three possibilities). The tariff values were between 11.49 and 87.25 (in million US$).

REMARKS AND CONCLUSIONS

Our detailed analysis of the frequency and severity of cyber risk considering two ways of approaching the ratemaking process for this type of risk showed how much the inclusion of covariates can increase the financial need to be charged as well as how much a premium value may change depending on the risk class.

According to our calculations, insurance premiums can become expensive, which could generate disinterest on the part of both insurers in accepting such a risk and policyholders due to the high cost.

Thank you! Obrigado!

Questions?

Contacts:
E-mail: alana.azevedo@ufc.br
Phone number: +55 85 999248203