CALCULATION MODELS OF NET RISK PREMIUMS FOR RCA INSURANCE CONTRACTS

Adriana NĂSTASE (DUMITRACHE), PhD student,
Romanian Academy,
Doctoral Department: Economic, Social and Legal Sciences
nstsadriana@yahoo.com

DOI: https://doi.org/10.36004/nier.cecg.III.2023.17.34

Abstract*. Starting from the fundamental role of insurance, that of providing protection by the insurer to the person interested in concluding an insurance contract (the insured) in exchange for the insurance premium (insurance price), advanced mathematical models have been developed for the appropriate determination of the insurance price. In Romania, RCA insurance is one of the most regulated insurances due to the importance it occupies from the point of view of the volume of gross premiums subscribed. In the total of gross premiums written in Romania, both general and life insurance, RCA insurance holds on average in the period 2014-2022 approximately 47.5%, hence the motivation for this research topic in order to establish the most appropriate premium for the insured RCA portfolio. In the case of non-life insurance, the estimation of the pure insurance premium is carried out with the help of generalized linear models (GLM). The segmentation of the insured portfolio into homogeneous risk classes is more than necessary for choosing the optimal explanatory variables, which reproduce the behavior of the response variable as accurately as possible.*
Keywords*: insurance, generalized linear model, distribution*
JEL: *G22, G52, I13, J65*
UDC: *330.4:364*

Introduction. General insurance faces many challenges, from fierce competition in the market or the evolution of the distribution channel used by consumers to the evolution of the regulatory environment. Price is the central link between solvency, profitability and market shares (volume). Taking into account the volatile nature of the risk covered by insurance products, an econometric substantiation of insurance product rates is needed. Thus, econometric pricing models can quantify the history observed by insurance companies as accurately as possible and ensure adequate prices for insurance products that provide stability and profitability. Through such econometric models developed over time, it has been empirically demonstrated that the solvency risk of insurance companies has decreased considerably and there have been fewer cases of insurance company bankruptcies.

Literature review. From the point of view of the distribution followed by the random variable the number of claims produced by some policy, through studies such as (Denuit et al, 2007), (De Jong & Heller, 2008), (Johansson & Ohlsson, 2010) on the portfolios analyzed by them, it was observed that this

variable follows a distribution from the exponential family among the following: Poisson, NegativeBinomial, Binomial and among these we note that the most often used is the Poisson distribution.

Although the behavior of drivers may be different from one country to another (here including the legislation), I believe that it cannot be significantly different because human nature and human behavior remain close regardless of the area, religion or culture.

To estimate the severity (Tevet et al, 2016) proposes a GLM model with 2 impulse variables, age and married or single status of the insured person. The authors include 976 policies in the analysis and use a Gamma distribution and a logarithmic link, and the results indicate that with increasing age, the cost of claims increases and unmarried people produce more claims. The authors (Tevet et al, 2016) find that an unmarried person produces claims costs about 16.18% more than married people.

According to the results of (Tevet et al, 2016) regarding age as an impulse variable are somewhat unexpected, because the increase from one age to another in the cost of claims reaches very high values, for example an insured aged 45 of years, produces 448.191% higher damages than a 30-year-old insured. However, this hypothesis cannot be generalized and must be adapted according to the specifics of the country and the policyholders in the insurer's portfolio. Both (Anderson et al, 2007) and (Tevet et al, 2016) the importance of analyzing the data prior to inclusion in the model and emphasize the identification of "outliers", which significantly distort the results of the coefficients. For example, (Tevet et al, 2016) proposes the use of the "Cook's distance" technique to identify the region of extreme values and exclude them before estimating the coefficients.

At the same time, the amount of damage of these extreme values should not be completely excluded, as a load percentage is used, applied after determining the premium through the GLM model. This topic is also carefully treated by (Denuit et al, 2007), who propose the segmentation of the total premium into 2 components, one component is the premium for normal (non-extreme) damages, and the second is the premium for extreme damages, both being estimated separately, and the total premium is their sum.

Research methodology. Calculation models of net risk premiums for RCA insurance contracts. The research study is based on the GLM econometric models for determining the risk premium related to RCA insurance contracts and for this, 2 distinct econometric models were used for PF and PJ, but also distinct econometric models for estimating the Frequency (probability of risk occurrence) and for Severity estimation (average damage). Following the analysis of damage distributions, the use of a Poisson distribution for Damage Frequency (both PF and PJ) and Gamma distribution for Damage Severity (both PF and PJ) was considered. Thus we can emphasize again the importance of GLM econometric models, which compared to the classic linear models where the error distribution is normal, can use a multitude of other distributions, which are more appropriate to the insurance field.

In the research study we showed that the distributions: Normal (Gaussian), Gamma, Inverse Gaussian, Poisson, Negative Binomial, Binomial, are part of the exponential family, these being the most present in the insurance pricing processes

(except for normal) . For cost modeling of non-life and especially motor insurance claims, the Gamma and Inverse Gaussian distributions are often used (also the LogNormal distribution is suitable and often used for cost modeling, but it is not part of the exponential family and requires models complex for parameter estimation). To model the number of accidents produced by a certain policy within the portfolio, the Poisson, Negative Binomial, Binomial distributions are used, being often used in specialty practice.

Thus, to estimate the frequency, the following explanatory variables (exogenous) are used with the following segmentation levels: age (18-24, 25-30, 31-60), gender (M, F), age of the car (3-5, <> 3-5), the type of fuel (gasoline, diesel), the type of locality (urban, rural), the power of the car expressed in Kw.

After estimating the coefficients for the frequency estimation model, using the Negative Binomial distribution, the results indicate the highest frequencies found for 18-24 female persons with more than 110 Kw and who also have gasoline fuel of approximately 46.82% . The final premium is given by frequency * average cost of claims (severity), and for this segment the authors estimated a premium of approximately 800 monetary units, being the highest within the portfolio.

The database in the initial form in which it was collected from the respective insurance company refers to a database that only includes passenger cars (maximum mass < 3500 kg and a maximum of 9 seats inclusive) that includes both natural persons (noted PF) and and legal entities (noted PJ). From this, only insurance contracts with an insurance contract duration of 12 months were selected, considering that they have a specific risk profile compared to those of other durations. It is recommended to proceed in this way and for the rest of the policies to calculate an average % deviation from the 12-month premium and thus, the insurer applies the resulting % to the 12-month premium. Illustrative example: premium result for 12 months = 1200 lei => premium for one month = 1200/ 12 * 1.5 = 150 lei, that 1.5 representing an additional adjustment coefficient due to potential differences in frequency and severity for this category of insured. The following Table shows the selection of the database from the initial contracts and those that remained to be used in the GLM pricing model.

Table no. 2: General database information separately for PF and PJ

Insuresd person type	Total No. of Contracts	12 months Contracts	Total exposure	Exposure related to the 12 months Contracts
PF	410029	338373	353496	332284
PJ	354761	293139	306583	288428
TOTAL	764790	631512	660079	620712

Source: own processing

Thus, from the total of 764,790 contracts (PF + PJ), only those with a contractual duration of 12 months were considered in the calculation, respectively a number of 631,512 contracts. Furthermore, checks were carried out regarding the informational quality of the 631,512 contracts, respectively:

- In RStudio, the following reasoning was applied "test_duplicates <- input_db_autoturisme %>% group_by(nrcontract) %>% summarize(rows= n())" and a number of 3 contracts were identified that are duplicates, not being material related to the total number;
- RStudio was used and the minimum and maximum age was checked for both PF and PJ and it was observed that for PJ it is 0 in both cases, which is normal, not being a criterion for PJ, while for PF the minimum age was found to be 0 years and maximum age of 106 years. PF cases under the age of 18 refer to 80 contracts, i.e. less than 0.1%;
- It was analyzed whether the cars comply with the primary condition, up to 3500 kg maximum mass and maximum 9 seats. The following were identified: 301 PF contracts and 1489 PJ contracts that do not comply with the mentioned conditions, representing 0.3% of the total contracts and having a damage volume of 940,000 lei. These have been excluded from the calculation and a uniform end error factor will be applied.
- Contracts were checked for complete information on: region, car make and if not identified, given the car make and/or region with the highest exposure.
- The recorded claims have been checked to ensure that they are not negative values. No negative claims volume or negative claims number were recorded for any contract.

In conclusion, no major data quality errors were recorded and the database is suitable for use in the GLM calculation model. Regarding the maximum authorized mass of up to 3500 kg and the maximum number of 9 seats, contracts that do not comply with these conditions were excluded and at the end of the results obtained from the GLM model, a uniform error factor will be applied for all homogeneous groups. Next, the distribution of the distinct claim volume for PF and PJ was analyzed and the extreme claims (outliers) that can distort the results of the GLM model and induce volatility in the estimators were limited.

Table no. 3: Top 10 PF damage and top 10 PJ damage, high damage limitation

Rank	top10_PF	top10_PJ
1	426,197	2,491,428
2	334,643	1,581,157
3	282,463	993,959
4	209,159	312,912
5	207,722	250,693
6	192,269	231,948
7	175,201	230,619
8	161,258	222,959
9	147,427	216,370
10	137,010	194,341
Total top 10	2,273,348	6,726,387
Total claims volume	72,308,735	103,745,802
Outlier weight	2.02%	4.91%
Top 10 weight	3.14%	6.65%

ource: own processing

After analyzing the top 10 claims in the case of PF and PJ, large damages were considered in PF over 200,000 RON, while in PJ over 330,000 RON, it should be noted that the share of outliers in PF is 2% and in PJ of 4.91%. Thus, these damages without exclusion induce volatility in the GLM model and distort the predictive power of the estimators.

Claims considered outliers (as described above) were excluded from the analysis, but the amount of claim was evenly distributed among the rest of the claims contracts, so that overall the amount of damage remained unchanged. Additionally, they were from the damage analysis under 300 RON, both for PF and for PJ, because they are too small and reduce the predictive power of the GLM model. Their number was relatively small, respectively 5 claims in the case of PF and 3 claims in the case of PJ.

The adjustment for past inflation was applied after excluding extreme claims (outliers) and the consumer price index according to INSE was used, alternatively depending on the availability of data at the time of the analysis, an inflation index related to the auto components sector can be used. The estimation by the Poisson method indicates small deviations from the Empirical Frequency, so in the case of PF the Poisson distribution represents the main distribution that will be used in the GLM. Next, the frequency distribution in the case of PJ is presented:

Table no. 4: Frequency distribution PJ

Claims number	Exposure	Empirical frequency	Poisson	Difference
0	274,838	95.3%	94.4%	-0.9%
1	11,566	4.0%	5.4%	1.4%
2	1,203	0.4%	0.2%	-0.3%
3	748	0.3%	0.0%	-0.3%
4	23	0.0%	0.0%	0.0%
5	6	0.0%	0.0%	0.0%
6	13	0.0%	0.0%	0.0%
7	12	0.0%	0.0%	0.0%
8	7	0.0%	0.0%	0.0%
9	2	0.0%	0.0%	0.0%
10	4	0.0%	0.0%	0.0%
Total	288,422	100.0%	100.0%	0.0%

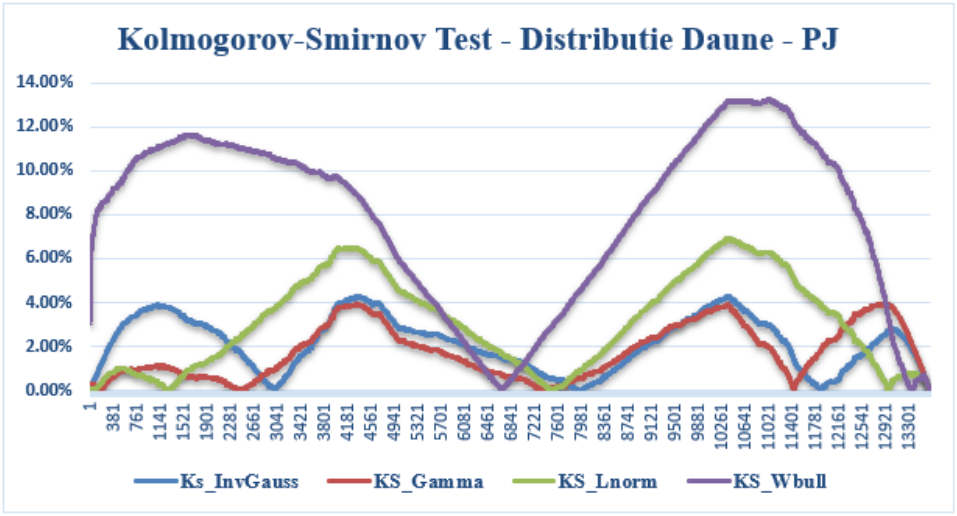
Source: own processing

To choose the distribution followed by damage, the Kolmogorov-Smirnov test was used, which is calculated as the modulus of the difference between the empirical (observed) Cumulative Distribution Function (CDF) and the distribution assumed to estimate the data. The theoretical distributions tested are: Gamma, LogNormal, Weibull, Inverse Gaussian, these were tested individually for both PF and PJ. The criterion for choosing the distribution by the Kolmogorov-Smirnov test involves choosing based on the difference between the minimum and the maximum

difference. Thus, for each value in the sample, the empirical_CDF is compared with the theoretical_CDF.

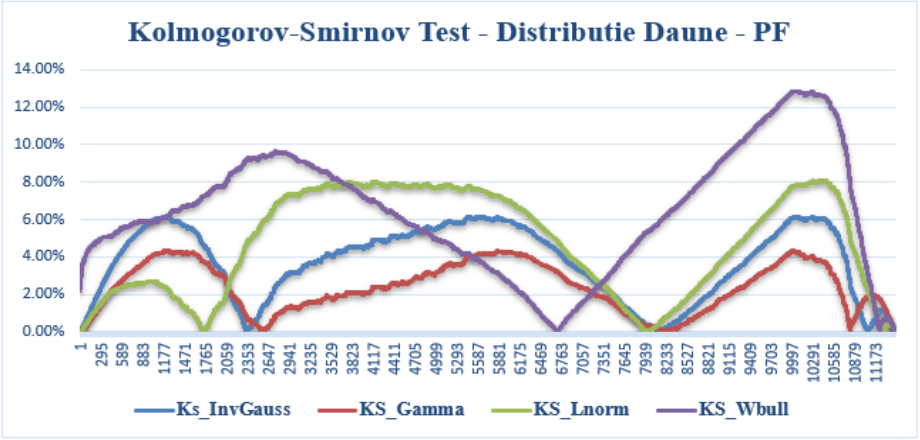
Graphical comparison for the KS test, where you can see the area with the largest deviations of the theoretical distribution (the one chosen by us) with the empirical one (result from the sample). Within the graph, on the X-axis, 1 represents the smallest damage and increases from 1 in 1 to the largest recorded damage, thus it is observed on the area of values where the KS test indicates the largest or smallest errors.

Chart no. 1: Comparative graphical visualization of the Kolmogorov-Smirnov test (noted in the KS paper) for PJ damages (legalentities)



Source: own processing

Chart no. 2: Comparative graphical visualization of the Kolmogorov-Smirnov test (noted in the KS paper) for PF damages (individuals)



Source: own processing

Main results. We demonstrated in the paper that the introduction of the geographic region and the car brand are essential factors in determining an appropriate RCA insurance premium using GLM models. The geographical region in view of the fact that it has been observed both statistically and in specialized literature that both the Frequency (probability of occurrence of the insured risk) and the Severity (average damage) differ according to the geographical region due to the road infrastructure specific to the area, the number of inhabitants (large cities have a higher frequency of accidents, but the damage caused is lower) and the infrastructure of shared public means that alternative travel measures compared to motor vehicles (insured persons drive less often and thus reduce the number of accidents produced in an insurance year).

The same thing was also confirmed within the car brand, namely that the homogeneous risk group where differences in Frequency and Severity were observed between car brands, for example, BMW brand cars were observed to have an average PF Frequency of 11% compared to the overall average of 4.5%. The same thing is observed within PJ (legal entities). Thus, the inclusion of the geographic region and the car brand in the GLM calculation model within the paper brought a significant improvement in the individualization of the risk and the appropriate setting of the premium according to the customer profile. It should also be noted that there is no work published with similar studies on insurance companies in Romania, this work representing a good guideline for insurance companies in Romania and what elements could be included or excluded from the calculation of RCA rates.

Conclusions. In conclusion, for estimating the net premium in general insurance, the use of generalized linear models (GLM) is more than necessary, and especially in auto liability insurance. Since 1972 and until today, generalized linear models have continued to be refined in order to be able to estimate as precisely as possible the behavior of the endogenous variable (response, denoted Y) according to the exogenous variables (explanatory, denoted Xi). The original elements of the present calculation model GLM cars PF and PJ for establishing the RCA pure risk premium, the fact that the RStudio program which is free was used throughout the calculation process would represent a great advantage for smaller insurance companies that do not want to invest heavily in specialized software for calculating GLM models. Automatically by purchasing such software increases the company's cost rate leading to higher insurance premiums that risk being no longer competitive and attractive to customers. Thus, although the insurance premium would increase (because a higher cost rate is included in the calculation formula), the total volume could decrease because customers are less attracted to that insurance product. In the case of RCA insurance products it would mean a lower risk dispersion and less liquidity which could have significant effects on the company's profitability, even bankruptcy in some cases and thus, there must be a balance between the insurance price and the client's expectations regarding these prices. Thus, the development of the entire process of calculating the GLM risk premiums for RCA through a statistical software that does not attract additional costs (in the present case RStudio) represents a great advantage that can be exploited by insurance companies.

REFERENCES

Anderson, D., Feldblum, S., Modlin, C., Schirmacher, D., Schirmacher, E., & Thandi, N. (2007). *A Practitioner's Guide to Generalized Linear Models*. Virginia: Towers Watson.

De Jong, P. & Heller, G. Z. (2008). *Generalized Linear Models for Insurance Data*. Cambridge: Cambridge University Press.

Denuit, M., Maréchal, X., Pitrebois, S. & Walhin, J. F. (2007). *Actuarial Modelling of Claim Counts*. Bruxells: John Wiley, Sons. https://doi.org/10.1002/9780470517420

Johansson, B., & Ohlsson, E. (2010). *Non-Life Insurnace Pricing with Generalized Linear Models*. first ed. Stockholm: Springer-Verlag, Berlin, Heidelberg.

Tevet, Dan, et al. (2016). *Generalized linear models for insurance rating*. Virginia: Casualty Actuarial Society.