STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **13**, June 2025, pp 2294–2315.
Published online in International Academic Press (www.IAPress.org)

Optimizing Automobile Insurance Pricing: A Generalized Linear Model Approach to Claim Frequency and Severity

Mekdad Slime [1,*], Abdellah Ould Khal [1], Abdelhak Zoglat [1], Mohammed El Kamli [2], Brahim Batti [1]

[1] *Laboratory of Mathematical, Statistics and Application (LMSA), Faculty of Sciences, Mohammed V University in Rabat, Morocco*
[2] *Laboratory of Economic Analysis and Modelling (LEAM), Faculty of Sciences, Economic, Juridical and Social-Souissi, Mohammed V University in Rabat, Morocco*

Abstract Morocco's insurance sector, particularly auto insurance, is experiencing significant growth despite economic challenges. To remain competitive, companies must innovate and adjust their pricing to meet customer expectations and strengthen their market position. Traditionally, actuaries have used the linear model to assess the impact of explanatory variables on the frequency and severity of claims. However, this model has limitations that do not always accurately reflect the reality of claims or costs, especially in auto insurance. Our study adopted the generalized linear model (GLM) to address these shortcomings, enabling a more precise statistical analysis that better aligns with market realities. This paper examines the application of GLM to model the total claim burden of an automobile portfolio and establish an optimal rate. The steps include data processing and analysis, segmentation of rating variables, and selecting appropriate distributions using statistical tests such as the Wald test and the deviance test, all performed using SAS software.

Keywords Financial and Insurance Mathematics, Auto Insurance, Generalized Linear Model (GLM), Actuarial Science, Applications of Statistics to Economics, Pricing

AMS 2010 subject classifications 91Gxx, 62J12, 62P05, 62P20

DOI: 10.19139/soic-2310-5070-2157

1. Introduction

Traditional linear regression models assume that the dependent variable follows a normal distribution and that the relationship between the dependent variable and the independent variables is linear. However, these assumptions are not suitable for many types of data, such as binary results or counts. Generalized Linear Models (GLMs) were developed to overcome these limitations by accommodating different types of distribution (e.g., Binomial, Poisson) and incorporating a link function that relates the mean of the dependent variable to the linear predictors, thereby transforming a non-linear relationship into a linear one. The concept of GLM was initially proposed by John Nelder and Robert Wedderburn in their seminal paper "Generalized Linear Models" [1]. Their work extended the linear modeling framework to encompass a variety of distributions beyond the normal distribution, which was the primary focus of classical linear regression. GLMs have been widely used in various fields such as economics [2, 3, 4], medicine [5, 6, 7], social sciences [7, 8, 9], biology [7, 10], and engineering [7, 11, 12], due to their flexibility.

Insurance is a mechanism through which the insurer commits to paying a benefit in the event of a predefined uncertain event, in exchange for a premium paid by the insured, as seen in automobile insurance. Insurance and reinsurance companies must be subject to specific oversight to ensure their social impact and preserve their crucial

*Correspondence to: Mekdad Slime (Email: slime.mekdad@gmail.com). Laboratory of Mathematical, Statistics and Application (LMSA), Faculty of Sciences, Mohammed V University in Rabat, B.P. 1014 Rabat, Morocco.

ISSN 2310-5070 (online) ISSN 2311-004X (print)

role in the financing of the economy [13, 14, 15, 16, 17]. In Morocco, this responsibility lies with the Insurance and Social Welfare Authority (ACAPS = Autorité de contrôle des assurances et de la prévoyance sociale) [18], which regulates the issuance of licenses for market entry and ensures continuous supervision. According to its annual statistics, in 2021, the premiums issued by insurance and reinsurance companies reached significant amounts, with 22.9 billion dirhams for life premiums, including acceptances, and 27.3 billion dirhams for non-life premiums. Such data highlight the central role played by insurance companies. Their impact extends beyond the social realm, contributing to well-being and compensating for third-party losses. They also emerge as essential contributors to the country's economic growth [19].

In insurance, the production cycle operates in reverse: the premium (the contribution from the insured) is collected upfront, while the indemnity payment (obligation of the insurer) occurs later. Consequently, the exact cost cannot be determined in advance. To address this, the insurer must take specific steps: selecting risks to minimize exposure to unfavorable outcomes, distributing risks through co-insurance and reinsurance to protect against significant losses, and pricing risks based on historical experience, employing statistical methods, probability calculations, and financial mathematics to ensure fair allocation of costs among insured [20, 21, 22, 23, 24]. To achieve these objectives, insurers increasingly rely on GLMs as a powerful risk assessment and pricing tool. GLMs allow insurers to model the relationship between various risk factors such as age, driving history, vehicle type, and the probability of claims, allowing more accurate predictions of expected losses. Using GLMs, insurers can refine their risk selection process, ensuring that premiums are closely aligned with the actual risk profile of each policyholder. Moreover, GLMs improve the ability to price risks in a manner that is both fair and financially sustainable, as they account for the complex, non-linear interactions between different variables. This statistical approach helps set fair premiums and supports the strategic management of risk distribution through co-insurance and reinsurance, ultimately contributing to the insurer's financial stability [7, 25, 26, 27].

Generalized Linear Models have recently attracted significant attention, particularly within the insurance sector, where researchers have extensively explored their applications in risk assessment, pricing, and maintaining insurance stability. In 2014, Kafková, S. et al. [28] utilized GLMs to predict annual vehicle insurance claim frequency based on data from 57,410 vehicles. Their approach involved evaluating models by comparing predictor variables through deviance analysis and the Akaike Information Criterion (AIC) with R software computations. Afterward, in 2015, David, M. [29] provided an overview of GLM techniques for calculating the pure premium based on observable policyholder characteristics. This work included a numerical illustration using a French auto insurance portfolio performed with SAS statistical software.

Subsequent studies further demonstrated the versatility of GLMs in the insurance industry. In 2018, Xie, S. et al. [30] offered general guidelines for applying predictive modeling to regulate insurance rates, highlighting GLMs' effectiveness in reviewing auto insurance rate filings. Following this, in 2019, Erik Šoltés et al. [31] published a paper focusing on claim severity analysis in motor third-party liability insurance using GLM based on anonymized data from a Slovak insurance company. Continuing this trajectory, in 2022, E. Seyam et al. [32] proposed alternative tariff systems for estimating pure premiums for Misr Insurance Company by employing GLM, Generalized Linear Mixed Models (GLMM), and Generalized Additive Models (GAM), using Gamma and Poisson distributions on a dataset comprising 576,381 insurance contracts from 2013 to 2016.

The application of GLMs in insurance has seen further refinement in recent years. In 2023, R Oktavia et al. [33] assessed Poisson and Negative Binomial GLM models for estimating auto insurance claim frequencies using R software. They discovered that while both models performed well for the Swedish dataset (dataOhlsson), they were unsuitable for the Australian dataset (ausprivauto0405). Most recently, in 2024, Esmeralda Brati [34] utilized GLM and Generalized Additive Models for Location, Scale, and Shape (GAMLSS) to estimate claim amounts, analyzing data from 229 Automobile Bodily Injury Claims provided by an Albanian insurance company. This study also explored the effects of various explanatory variables on claim amounts, further emphasizing the broad applicability of GLMs in the insurance sector.

Our research applies GLMs to model the total claim burden of an automobile insurance portfolio, providing a structured approach to overcoming limitations in traditional pricing methods. Automobile insurance plays a vital role in the Moroccan insurance market, accounting for 26.1% of non-life premiums in 2021. Traditional approaches

often struggle to capture the intricate relationships in claims data, whereas GLMs offer the flexibility to model non-linear interactions and accommodate diverse data distributions.
This paper makes several key contributions, including developing a practical framework for calculating fair and sustainable premiums. It employs advanced segmentation techniques to create homogeneous risk classes and selects optimal statistical distributions for modeling claim frequency and severity. The methodology integrates thorough data cleaning, exploratory analysis, and rigorous statistical testing, utilizing the Negative Binomial and Gamma distributions. By leveraging real-world data from the Moroccan market, our study delivers actionable insights for insurers seeking to enhance pricing accuracy, ensure regulatory compliance, and strengthen competitive positioning. In addition to demonstrating the value of GLMs in modern actuarial science, this work lays the groundwork for future innovations in data-driven risk management and premium optimization.

The remainder of this paper is structured as follows. First, the structure and evolution of premiums and market shares in the Moroccan insurance market from 2019 to 2021 are discussed. Next, the theoretical framework for automobile insurance pricing, emphasizing GLM methods, is presented. Following this, the application of these theoretical concepts is detailed, along with a discussion of the resulting findings.

2. Global market vision

Tables 1 and 2 show the distribution of premiums issued in 2021 and the market shares of the leading companies, respectively. Figure 1 depicts the evolution of premiums issued from 2019 to 2021, differentiating between non-life insurance and life insurance & capitalization.

Table 1. The structure of premiums issued for the year 2021 and the contribution of each branch on the Moroccan insurance market [35].

Branches	**Primes**	**Contribution**
Life Insurance and Capitalization	22 942,30	46,1%
Automobile	12 988,70	26,1%
Personal Accidents	4 772,80	9,6%
Worker's Compensation	2 319,60	4,7%
Fire	2 053,60	4,1%
Assistance	1 513,90	3,0%
Other Non-Life Policies	888,20	1,8%
Transport	706,40	1,4%
General Third-Party Liability	604,40	1,2%
Guarantee against the consequences of catastrophic events	521,80	1,0%
Technical Risks	276,10	0,6%
Reinsurance Acceptance	221,70	0,4%
Total (in millions of dirhams)	49 809,50	100,0%

Remark 2.1
Table 1 highlights the following points:

- In 2021, life insurance & capitalization accounted for 46.1% of premiums issued, while other insurance branches represented 53.9%.
- Auto insurance dominated the non-life insurance market with an annual turnover of 12.99 billion dirhams in 2021, followed by personal accident insurance with 4.77 billion dirhams, and worker's compensation insurance with 2.32 billion dirhams.
- The prominence of auto insurance is largely attributed to mandatory civil liability insurance and the tendency of cautious policyholders to seek extensive protection against common risks.

Table 2. The evolution and market shares of the main companies [35].

Compagnie	2019	2020	2021	Evolution 2020/2021	Market Share
Wafa Assurance	8 853,0	8 374,2	9 088,9	8,5%	18,2%
RMA	6 816,0	6 876,0	7 680,7	11,7%	15,4%
Mutuelle Taamine Chaabi	5 123,2	5 787,3	6 308,4	9,0%	12,7%
Sanlam Assurance	5 422,4	5 126,0	5 621,1	9,7%	11,3%
Axa Assurance Maroc	4 645,2	4 871,7	5 567,4	14,3%	11,2%
Atlanta Sanad	4 840,8	4 937,6	5 400,8	9,4%	10,8%
Marocaine Vie	2 267,6	2 158,2	2 339,1	8,4%	4,7%
MCMA	1 541,2	1 798,0	2 067,0	15,0%	4,1%
Allianz Assurance Maroc	1 479,9	1 572,3	1 426,8	-9,3%	2,9%
MAMDA	1 034,6	1 092,5	1 172,1	7,3%	2,4%
CAT	693,0	694,1	779,8	12,3%	1,6%
MATU	416,6	525,7	714,1	35,8%	1,4%
Maroc Assistance Internationale	568,1	561,5	620,9	10,6%	1,2%
Africa First Assist	471,7	325,8	322,0	-1,2%	0,6%
Wafa Ima Assistance	281,4	258,3	281,6	9,0%	0,6%
Euler Hermes ACMAR	144,9	136,3	134,2	-1,5%	0,3%
RMA Assistance	113,1	109,2	123,0	12,6%	0,2%
Coface Maroc	62,7	81,1	71,2	-12,2%	0,1%
AXA Assistance Maroc	86,9	47,0	55,8	18,7%	0,1%
Smaex	39,9	27,6	34,8	26,1%	0,1%
Total (in millions of dirhams)	**44 902,2**	**45 360,4**	**49 809,7**	**9,8%**	**100,0%**

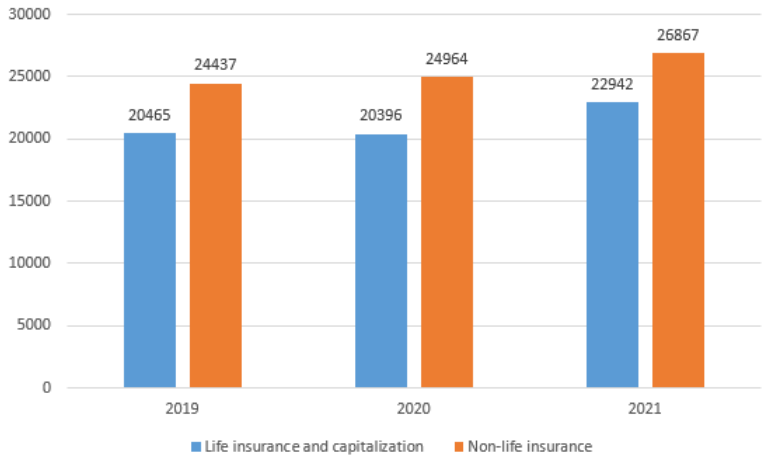

Figure 1. Evolution of premiums issued 2019-2021 [35].

Remark 2.2

1. Table 2 shows that the top six companies hold 79.6% of the life and non-life insurance market. Wafa Assurance leads with 18.2%, followed by RMA (15.4%), Mutuelle Taamine Chaabi (12.7%), Sanlam Assurance (11.3%), Axa Assurance Maroc (11.2%), and Atlanta Sanad (10.8%).
2. Figure 1 indicates a modest 1% increase in premiums issued in 2020 compared to 2019, reflecting resilience amidst the COVID-19 pandemic despite a generally slow growth period. In 2021, there was a significant recovery, with premiums rising by nearly 10%, driven by renewed activity in both non-life and life insurance sectors. In the automobile sector, premiums saw a slight increase of 0.1% in 2020, followed by a more substantial rise of 8.6% in 2021.

3. Materials and methods

3.1. Principles of Pricing for Civil Liability Insurance

In the insurance market, insurers work to maintain profitability and competitiveness by regularly introducing new products to meet evolving market demands. The product development process encompasses several stages that require close coordination among different departments and stakeholders. One critical stage is pricing, handled by the actuarial department, which involves determining the appropriate premium levels for policyholders.

3.1.1. The frequency-cost theory of claims

The insurance operation involves transferring the risk, either fully or partially, from the insured to the insurer. The primary challenge for insurers is to accurately evaluate and control the associated risks. This process includes estimating the expected claims burden, known as the pure premium, which is the amount needed to cover claims without accounting for administrative costs. The pure premium reflects the inherent risk to the insurer. In a priori pricing, the goal is to set a premium that closely matches this expected claims burden. For a given policy and period, the desired premium is the pure premium, determined by minimizing the deviation from the anticipated claims: $\min_{P>0} d(P,S) := \mathbb{E}(S-P)^2$. After a simple calculation, we find that $P = \mathbb{E}(S)$. Therefore, pricing consists of modeling the claims.

In insurance, costs for a portfolio of individual risks can be modeled using different approaches, primarily categorized into individual risk models and collective risk models. We will focus on the collective risk model, which is essential in auto insurance. In this model, the total claims burden S follows a compound distribution, defined by $S = \sum_{i=1}^{N} C_i$, with:

- N represents the number or frequency of claims observed during the insurance period.
- C_i are independent and identically distributed random variables representing the amounts or costs of the i-th claim, without considering the specific insured individual.
- N and C_i are independent (Cost-Frequency independence).

Remark 3.1

The collective model reveals that two essential factors influence the distribution of S: the frequency of claims and their costs. In other words, it is more relevant to model the frequency and costs of claims separately rather than modeling the total burden directly. Under the assumption of independence between N and C_i, the following property arises from the collective model: $\mathbb{E}(S) = \mathbb{E}(N) \times \mathbb{E}(C)$. This equation shows that the desired pure premium equals the average claim frequency multiplied by the average cost per claim.

Remark 3.1 highlights that the pure premium formula depends on a mathematical expectation, which requires a large portfolio to estimate the average claims burden with precision. In auto insurance, this condition is usually fulfilled due to mandatory liability coverage, allowing for pricing based on the Law of Large Numbers. Nevertheless, due to the heterogeneous nature of risks, the Law of Large Numbers does not apply directly to an insured portfolio. Consequently, insurers must divide the portfolio into homogeneous classes, ensuring that each class is sufficiently large to validate the model.

In summary, when setting rates, insurers must consider two key factors:

- A sufficient number of insured persons: Maintaining a sufficiently large portfolio of insured individuals.
- Homogeneity of risks: Creation of homogeneous classes within the portfolio of insured persons.

Remark 3.2

Based on the above, we will use conditional expectations to calculate the pure premium for a policyholder in the homogeneous class X. In other words, we will have the following formula:

$$\mathbb{E}(S|X) = \mathbb{E}(N|X) \times \mathbb{E}(C|X).$$

3.1.2. Segmentation

Segmentation entails dividing a portfolio into distinct, homogeneous subgroups, each comprising individuals with similar behaviors. The principle behind this is that each individual should pay a premium commensurate with their own risk. This method enables the evaluation and management of the consistency between premiums and claims within homogeneous risk groups, offering crucial technical insights. As a result, segmentation aids in implementing technical strategies at all levels, particularly in pricing and underwriting.

Remark 3.3

When performing segmentation, it is crucial to consider the condition of having numerous risks. Indeed, when aiming for an individual premium rate, the estimation of the pure premium based on the collective model can be biased, leading to poor results. Since the estimation of the average load will rely on a tariff class that is too narrow, it fails to leverage the Law of Large Numbers.

3.2. Generalized Linear Model (GLM)

Actuaries have long relied on the linear model to quantify the impact of explanatory variables on claim frequencies and severities. However, this model presents a series of limitations that do not always align with the reality of claim numbers or costs. Indeed, the linear model, defined in matrix form as $Y = X^T\beta + \varepsilon$ assumes that the mean is a linear function of the variables, which is not always the case, particularly in auto insurance. To overcome these restrictions the model has been expanded to include GLMs, representing a diverse set of statistical methods. GLMs extend the traditional linear model in two key ways: first, they accommodate a range of distributions, such as Normal, Poisson, and Gamma, rather than assuming a Normal distribution. Second, GLMs model a monotonic transformation of the mean as a linear function of the explanatory variables rather than modeling the mean directly. This approach relaxes both the normality assumption and the linearity requirement of the mean concerning the explanatory variables. What distinguishes GLM is that, instead of modeling the response variable Y, directly, a function of the expected value of this variable, known as the link function, is modeled. In other words, a mathematical transformation is applied to the expected value of the response variable, considering the actual distribution of errors.

3.2.1. The principle of GLM

A GLM model relates the expectation of Y to the explanatory variables as follows:

$$g(\mathbb{E}(Y_i)) = g(\mu_i) = \sum_{j=1}^{p} x_{ij}\beta_j.$$

with: $i \in \{1, 2, 3, ..., N\}$, Y_i is a random variable belongs to the exponential family, g is an invertible function, called a link function, and $\mathbb{E}(Y_i)$ depends on $\sum x_j\beta_j$ through the link function g.

3.2.2. The exponential family

The exponential family encompasses a set of probability distributions that can be expressed by a single general formula and share common properties. Distributions such as the normal, binomial, Poisson, and Gamma distributions all belong to this exponential family.

In all GLM models, we will assume that the response variable Y follows a density of the following form:

$$f(y|\theta, \phi) = exp\left[\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right].$$

with: a, b and c are known and differentiable functions, ϕ is the dispersion parameter, and θ is the natural parameter. It is linked to the first two moments of the law.

The following two formulas result from this: $\mathbb{E}(Y) = \mu = b'(\theta)$ and $Var(Y) = \sigma^2 = b''(\theta)a(\phi)$.

Depending on the type of response variable Y, a density function is selected from those listed in the table 3.

Table 3. Summary of the main factors of GLM by distribution.

Distribution	θ_i	$b(\theta_i)$	$a(\phi)$	μ_i
$\mathcal{B}(1,\pi_i)$	$\ln\left(\frac{\pi_i}{1-\pi_i}\right)$	$\ln(1+e^{\theta_i})$	1	$\pi_i = \frac{e^{\theta_i}}{1+e^{\theta_i}}$
$\mathcal{NB}(n,p)$ (n constant)	$\ln(p)$	$n\ln(1+e^{\theta_i})$	1	$e^{\theta_i} = \frac{n(1-p)}{p}$
$\Gamma(\mu_i,\nu)$	$-\frac{1}{\mu_i}$	$-\ln(-\theta_i)$	$\frac{1}{\nu}$	$-\frac{1}{\theta_i} = \mu_i$
Pois(λ_i)	$\ln(\lambda_i)$	e^{θ_i}	1	$e^{\theta_i} = \lambda_i$
$\mathcal{N}(\mu_i,\sigma^2)$	μ_i	$\frac{\theta_i^2}{2}$	σ^2	$\theta_i = \mu_i$

3.2.3. Link functions

Let g be a monotonic and differentiable link function. This function represents the relationship between the endogenous variable and the deterministic component. The common link functions are summarized in the table 4.

Table 4. Classic link functions

Identity function	$g : x \longrightarrow x$
Logarithm function	$g : x \longrightarrow \ln(x)$
Inverse function	$g : x \longrightarrow \frac{1}{x}$
Logit function	$g : x \longrightarrow \ln\left(\frac{x}{1-x}\right)$
Probit function	$g : x \longrightarrow \phi(x)$

Where ϕ is the distribution function of the normal distribution $\mathcal{N}(0,1)$.

3.2.4. WALD test

To assess the validity of a GLM, various criteria and tests are employed to ensure the model's relevance. The Wald test, in particular, is used to evaluate the significance of the explanatory variables. This test helps identify which variables should be retained and which ones significantly influence the response variable. Specifically, if a parameter β_j is zero, it suggests that the explanatory variable associated with the coefficient x_j does not affect the response variable Y. Therefore, testing whether a parameter is zero in the population under study is crucial.

The Wald test confronts the following hypotheses:

$$H_0 : \beta_j = 0 \text{ (there is no connection between } x_j \text{ and } Y)$$
$$H_1 : \beta_j \neq 0 \text{ (there is a link between } x_j \text{ and } Y)$$

4. Experimental results

4.1. Descriptive study

Before conducting any statistical analysis, it was crucial to thoroughly clean the databases. The collected data were often inaccurate, inconsistent, or redundant, which could distort the results. Therefore, we first presented the raw databases and then ensured data consistency by addressing outliers, missing values, and duplicates. This approach allowed us to perform a precise and reliable descriptive statistical analysis. This study focused on auto liability insurance contracts for tourism use, utilizing data from two files: a production file and a claims file. We began by importing the data from these files using the $proc\ import$ command in SAS. The production file contained 145 367 observations across 9 variables, while the claims file included 65 372 observations across 4 variables.

#	Variable	Type	Len	Format	Informat
1	Ex	Num	8	BEST12.	BEST32.
2	Com	Char	1	$1.	$1.
3	PF	Char	5	$5.	$5.
4	Sex	Char	1	$1.	$1.
5	exe	Num	8	BEST12.	BEST32.
6	DOB	Num	8	DDMMYY10.	DDMMYY10.
7	DMC	Num	8	DDMMYY10.	DDMMYY10.
8	n_police	Num	8	BEST12.	BEST32.
9	Zone	Char	22	$22.	$22.

#	Variable	Type	Len
1	n_police	Num	8
2	a_reference	Num	8
3	n_sinistre	Num	8
4	m_sinistre	Num	8

Figure 2. Lists of variables and their types from both files (SAS output).

We then analyzed the missing values, which included information either not reported by the insured or not recorded by the claims managers. Tables 5 and 6 summarize the number of missing values for each variable and their percentage relative to the entire database.

Table 5. The number of missing values (production file)

Observations	Variable name	Number of missing values	Percentage
1	Exhibition	0	0%
2	Combustion	15 038	10%
3	Fiscal Power	0	0%
4	Sex	10 222	7%
5	Exercise	0	0%
6	Date of birth	1 067	1%
7	Date of entry into circulation	15 038	10%
8	Police number	0	0%
9	Zone	0	0%

Table 6. The number of missing values (claims file)

Observations	Variable name	Number of missing values	Percentage
1	Police number	0	0%
2	Reference year	0	0%
3	Number of claims	0	0%
4	Amount of claims	0	0%

To address the missing values shown in Tables 5 and 6, and given the large size of our dataset, we chose to remove them. Consequently, we ended up with 125 105 observations in the production file and 65 372 observations in the claims file. We then moved on to the analysis of outliers. For the production file, we calculated the ages of drivers and vehicles, removing outliers by excluding vehicles aged less than 0 or greater than 80, as well as drivers aged less than 18 or greater than 80. After this correction, the production file contained 124 479 observations, with no other variables showing abnormal values. In the claims file, the analysis of outliers involved calculating the number of claims and the total amount. After removing negative amounts, the database consisted of 50 676 observations. Finally, we analyzed duplicates by using the SQL query "$select\ distinct^*$" in SAS. This analysis revealed no duplicates in either file. As a result, the number of observations in the production file remained at 124 479, while the number of observations in the claims file remained at 50 676, and the study relied on a final database, created by merging the production and claims files.

Remark 4.1
Our study relied on a final database, created by merging the production and claims files. As a result, each insurance contract was associated with all relevant variables included in our analysis. In other words, it was essential to combine the two existing databases.

4.1.1. Distribution of the automobile portfolio

This section categorizes the portfolio into four groups: sex, fiscal power, type of combustion, and circulation zone.

a- Distribution of automobile portfolio according to the sex of the insured

Sex or gender is a significant variable influencing the frequency of claims in auto insurance. Behavioral differences between males and females often result in distinct risk profiles. By analyzing the model coefficients associated with this variable, insurers can quantify these behavioral differences and adjust premiums accordingly to reflect the relative risk. Such insights are critical for ensuring fair and equitable pricing in auto insurance portfolios.

Sex	Frequency	Percent	Cumulative Frequency	Cumulative Percent
F	14752	11.70	14752	11.70
M	111300	88.30	126052	100.00
Frequency Missing = 30				

Figure 3. Proportion of males and females in the portfolio (SAS output).

In the insured population, men outnumber women. Specifically, men comprise 88.30% of the portfolio, while women represent 11.70%.

Sex	Frequency	Percent	Cumulative Frequency	Cumulative Percent
F	2705	18.58	2705	18.58
M	11856	81.42	14561	100.00
Frequency Missing = 7				

Figure 4. Proportion of males and females affected by disasters in the portfolio (SAS output).

Among the insured casualties, there are 11 856 men and 2 705 women. Men constitute 81.42% of the total casualties. Additionally, the percentage of men who are casualties relative to the total male population is 10.65% $(11856/111300 \times 100 = 10.65\%)$. In contrast, women who are casualties represent 18.33% of the total female population $(2705/14753 \times 100 = 18.33\%)$.

These findings suggest that while men represent a larger proportion of the insured population, women experience a higher casualty rate relative to their total population. This could indicate differences in driving behavior or exposure to risk between genders, which is an important factor for pricing models in the insurance industry.
By interpreting the model coefficients associated with gender, insurers can better understand how the sex of the insured influences both claim frequency and severity, allowing for more accurate risk assessments and premium adjustments.

b- Distribution of the automobile portfolio according to fiscal power

Fiscal power, a key variable associated with the insured vehicle, is a significant determinant in assessing auto insurance risk. Higher fiscal power often correlates with more powerful and faster vehicles, which can increase the likelihood of accidents due to higher speeds and performance capabilities. Vehicles with higher fiscal power are often driven more aggressively, further elevating the risk of claims related to both frequency and severity.

pf	Frequency	Percent	Cumulative Frequency	Cumulative Percent
10–14	17337	13.75	17337	13.75
7–10	74189	58.84	91526	72.59
<7	33042	26.21	124568	98.80
>=14	1514	1.20	126082	100.00

Figure 5. Distribution of the automobile portfolio according to fiscal power (SAS output).

According to the table in Figure 5, vehicles with less than 10 and more than 7 horsepower represent the majority, comprising nearly 60% of the portfolio. This finding suggests that the majority of insured vehicles fall within a mid-range of fiscal power, balancing between performance and potential risk. While these vehicles may not be as high-risk as those with significantly greater fiscal power, their frequency of claims may still reflect the risk associated with more powerful vehicles.
By interpreting the model coefficients for fiscal power, insurers can gain valuable insights into how changes in fiscal power influence risk, allowing for the development of more nuanced pricing models that better account for the correlation between vehicle performance and accident likelihood.

c- Distribution of the automobile portfolio according to the combustion type

The combustion type of a vehicle—specifically whether it is powered by Diesel or Gasoline—plays a crucial role in determining auto insurance risk. Diesel vehicles are often associated with higher fuel efficiency and durability, but they may also carry a higher risk of severe accidents due to their typically larger engine sizes and greater torque. Gasoline vehicles, on the other hand, are more commonly used in a wider variety of vehicle types and are generally considered to have lower environmental impact compared to diesel engines, though they can still pose significant risks depending on the vehicle's size and performance.

Com	Frequency	Percent	Cumulative Frequency	Cumulative Percent
G	38539	30.57	38539	30.57
D	87543	69.43	126082	100.00

Figure 6. Distribution of the automobile portfolio according to the combustion type (SAS output).

In our portfolio, 87 543 policies cover Diesel vehicles, representing 69.43%, while 38 539 policies cover Gasoline vehicles, accounting for 30.57%. This suggests that Diesel vehicles dominate our portfolio, likely due to their suitability for more frequent use and longer distances, despite their higher purchase cost.

Com	Frequency	Percent	Cumulative Frequency	Cumulative Percent
G	4018	27.58	4018	27.58
D	10550	72.42	14568	100.00

Figure 7. Distribution of damaged vehicles according to the combustion type (SAS output).

Among the insured vehicles, 4 018 Gasoline cars were involved in accidents, compared to 10 550 Diesel cars. The percentage of Gasoline cars involved in accidents relative to the total number of Gasoline vehicles is 10.4% $(4018/38539 \times 100 = 10.4\%)$. In contrast, the percentage of Diesel vehicles involved in accidents is 12% of the total Diesel vehicle population $(10550/87543 \times 100 = 12\%)$.
By interpreting the model coefficients for combustion type, insurers can evaluate the influence of fuel type

on accident frequency and severity, leading to more accurate pricing models that account for the distinct risks associated with diesel and gasoline vehicles.

d- Distribution of the automobile portfolio according to the circulation zone

The geographical area where a driver resides with their vehicle, referred to as the circulation zone, significantly impacts the accident rate and, consequently, the cost of auto insurance premiums. Areas with higher population density, such as urban centers, tend to experience higher traffic volumes, leading to an increased likelihood of accidents. This elevated risk results in higher insurance premiums for vehicles located in these zones. In contrast, rural or less populated areas often see fewer accidents, leading to relatively lower premiums.

Zone	Frequency	Percent	Cumulative Frequency	Cumulative Percent
BeniMellal-Khenifra	6252	4.96	6252	4.96
Casablanca-Settat	25432	20.17	31684	25.13
Dakhla-Oued Ed Dahab	6198	4.92	37882	30.05
Draa-Tafilalet	5144	4.08	43026	34.13
Fes-Meknes	17520	13.90	60546	48.02
Guelmim-Oued Noun	6373	5.05	66919	53.08
Laayoune-Sakia El Hamr	6245	4.95	73164	58.03
Marrakech-Safi	12403	9.84	85567	67.87
Oriental	8790	6.97	94357	74.84
Rabat-Sale-Kenitra	18885	14.98	113242	89.82
Souss-Massa	4984	3.95	118226	93.77
Tanger-Tetouan-Hoceima	7856	6.23	126082	100.00

Figure 8. Distribution of the automobile portfolio according to the circulation zone (SAS output).

Figure 8 shows that the regions of Casablanca-Settat and Rabat-Salé-Kénitra account for approximately 35% of the insured population.

By analyzing the model coefficients associated with circulation zones, insurers can gain a deeper understanding of how geographical factors influence claim frequency and severity. This insight enables more precise pricing models that reflect the risk variation based on the driver's location.

4.1.2. Mean of variables and correlation analysis

This section outlines the descriptive statistics of the variables and explores their correlations to assess relationships between the explanatory variables.

Simple Statistics						
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
n_police	126082	812200	491111	1.02404E11	3328	1632151
Ac	126082	45.36923	12.47993	5720243	18.00000	80.00000
Av	126082	12.69416	8.99844	1600505	0	76.00000
n_sinistre	126082	0.12443	0.35821	15688	0	4.00000
chargetot	126082	2993	18467	377349589	0	1099811
offset	126082	-0.92490	1.06640	-116613	-5.90263	0
freq	126082	0.52643	6.74371	66373	0	366.00000

Figure 9. Descriptive statistics of the variables (SAS output).

In our portfolio, the average claims frequency was 0.53, indicating roughly one claim for every two insured individuals. Additionally, the average driver was 45 years old, and the average vehicle was 12 years old (see Figure 9).

Pearson Correlation Coefficients, N = 126082
Prob > |r| under H0: Rho=0

	n_police	Ac	Av	n_sinistre	chargetot	offset	freq
n_police	1.00000	−0.02594 <.0001	0.01488 <.0001	−0.00247 0.3798	−0.00538 0.0563	−0.03113 <.0001	0.00038 0.8940
Ac	−0.02594 <.0001	1.00000	−0.05787 <.0001	0.02627 <.0001	0.00253 0.3693	0.13214 <.0001	−0.00615 0.0291
Av	0.01488 <.0001	−0.05787 <.0001	1.00000	−0.16047 <.0001	−0.04045 <.0001	−0.08717 <.0001	−0.03526 <.0001
n_sinistre	−0.00247 0.3798	0.02627 <.0001	−0.16047 <.0001	1.00000	0.44008 <.0001	0.11851 <.0001	0.22070 <.0001
chargetot	−0.00538 0.0563	0.00253 0.3693	−0.04045 <.0001	0.44008 <.0001	1.00000	0.04614 <.0001	0.10380 <.0001
offset	−0.03113 <.0001	0.13214 <.0001	−0.08717 <.0001	0.11851 <.0001	0.04614 <.0001	1.00000	−0.14657 <.0001
freq	0.00038 0.8940	−0.00615 0.0291	−0.03526 <.0001	0.22070 <.0001	0.10380 <.0001	−0.14657 <.0001	1.00000

Figure 10. Analysis of correlations between the explanatory variables (SAS output).

Referring to the table in Figure 10, we observe that the frequency of claims is correlated with the variables age of the driver and age of the vehicle (p-values are below 5%), suggesting that these variables can explain the frequency and should be retained in our modeling. Additionally, the age of the driver and the age of the vehicle are not correlated with each other, indicating their independence and confirming that neither variable will be removed. (As discussed in Article [19], Section IV, alternative methods can be employed to evaluate the relationship between these variables).

4.1.3. Adjustment of the total claims burden

In this section, we compare the distribution of the total load with several continuous probability distributions, namely the exponential distribution, the log-normal distribution, the Weibull distribution, and the Gamma distribution, using a Q-Q plot.

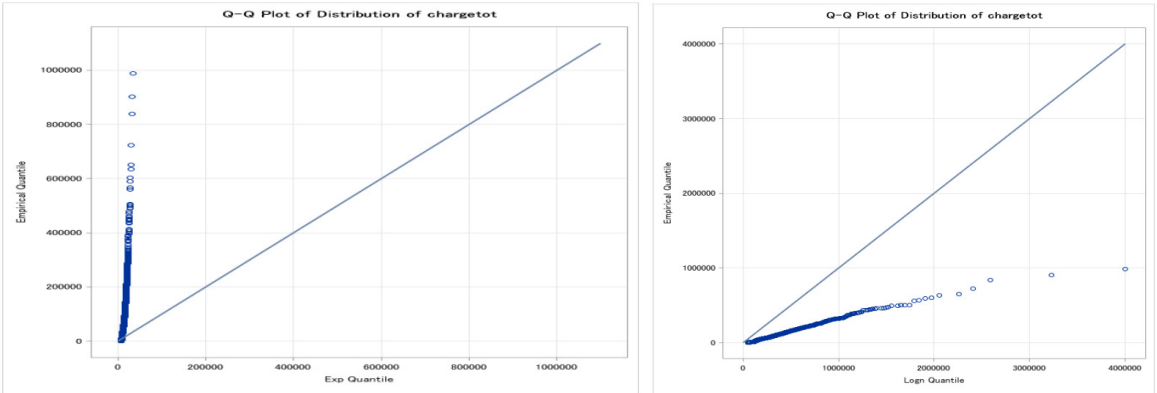

Figure 11. Adjustment the total claims burden to the laws: exponential (left) and log normal (right).

The Q-Q plot diagrams in Figure 11 reveal a significant deviation between the total load distribution and the exponential distribution and show that the total load curve is quite distant from the line of the log-normal distribution. This suggests that neither the exponential nor the log-normal distribution fits our data well.

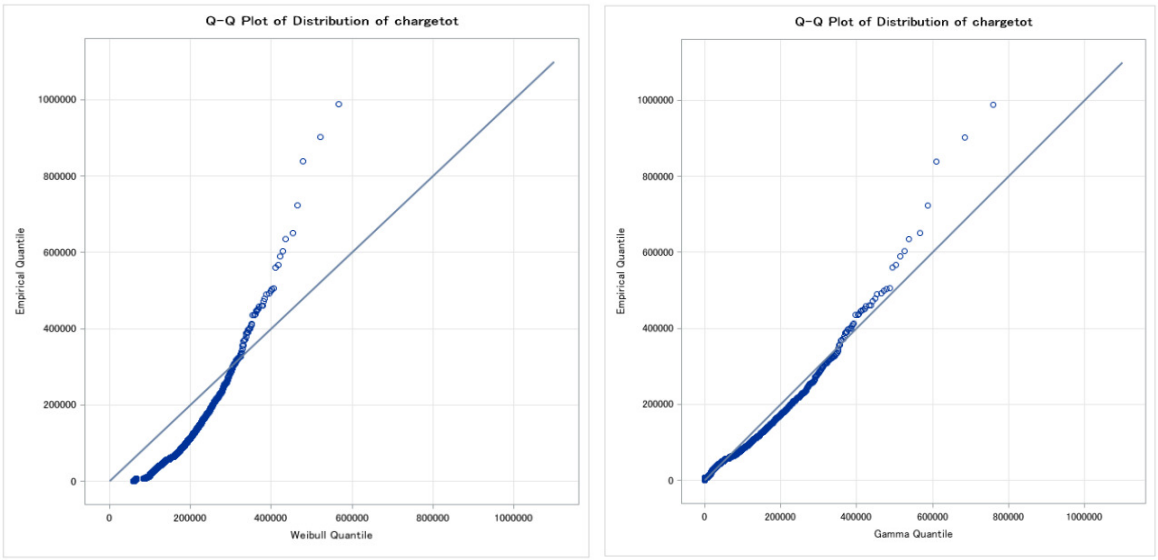

Figure 12. Adjustment of the total claim burden to the distributions: Weibull (left) and Gamma (right).

From the Q-Q plot above (Figure 12), it appears that the Weibull distribution provides a satisfactory fit to the total claims load distribution. However, the total load distribution is also quite close to the line of the Gamma distribution, suggesting that the Gamma distribution might be the most suitable for fitting our data.

4.2. Segmentation

Segmentation divides a portfolio into homogeneous risk classes where individuals with similar behaviors are grouped and pay the same premium. It ensures fairness for the insured and financial stability for the insurer by reducing adverse selection risk. To perform this segmentation, we have chosen the software ”SAS Enterprise Miner”, which assisted us in creating homogeneous classes of insured individuals. According to remark 3.1, two essential factors were involved: the frequency and the costs of claims.

4.2.1. Segmentation of tariff variables for modeling claims frequencies

We aim to define homogeneous risk classes with similar cost levels. The variables fall into two main categories: endogenous variables, such as the number of claims, and exogenous variables, which include factors related to the vehicle (age, fiscal power, etc.) and the driver (sex, age, etc.). Table 7 presents the various classes that have been created:

Table 7. The different classes built using ”SAS Enterprise Miner”

Variables	Classes
Vehicle age (av)	< 2.5
	between 2.5 and 10.5
	between 10.5 and 16.5
	> 16.5
Conductor age (ac)	< 30.5
	between 30.5 and 48.5
	> 48.5
Circulation zone (zone)	’Tanger-Tetouan-Hoceima’,’Rabat-Sale-Kenitra’,’Oriental’,’Guelmim-Oued Noun’,’Souss-Massa’
	other regions
Fiscal power (pf)	between 10 and 14
	others fiscals powers

The tables below (Figure 13) display the number of observations for each class created based on the previously mentioned variables.

av	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	18628	14.77	18628	14.77
2	40550	32.16	59178	46.94
3	22572	17.90	81750	64.84
4	44332	35.16	126082	100.00

ac	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	14904	11.82	14904	11.82
2	62277	49.39	77181	61.21
3	48901	38.79	126082	100.00

Figure 13. The frequency of observations in class ”vehicle age” (left) and class ”conductor age” (right).

The tables in Figure 13 show that vehicle age has been grouped into four classes with 18 628, 40 550, 22 572, and 44 332 observations, respectively. Driver age is divided into three classes with 14 904, 62 277, and 48 901 observations, respectively.

zone	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	46888	37.19	46888	37.19
2	79194	62.81	126082	100.00

pf	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	17337	13.75	17337	13.75
2	108745	86.25	126082	100.00

Figure 14. The frequency of observations in class ”circulation zone” (left) and in class ”fiscal power” (right).

The tables in Figure 14 show that the zone variable has been grouped into two classes with 46 888 and 79 194 observations, respectively. The fiscal power variable has also been grouped into two classes, with 17 337 and 108 745 observations, respectively.

4.2.2. Segmentation of tariff variables for average cost modeling

Before starting the segmentation of tariff variables explaining the average cost of claims, we analyzed the database. Using the ”$Proc\,Univariate$” procedure in SAS, we obtained the following table.

Basic Statistical Measures			
Location		Variability	
Mean	24740.80	Std Deviation	47275
Median	6000.00	Variance	2234909789
Mode	6000.00	Range	1089810
		Interquartile Range	23231

Figure 15. Descriptive statistics of the average cost base.

Based on the above statistics (Figure 15), the average of the ”average cost” variable was 24 740.8. In other words, a claim costs an average of 24 740.8. We also noted that the standard deviation of this variable was very high (47 275), which suggested the presence of severe claims. A common issue in non-life insurance is the significant impact of severe claims. When extreme amounts significantly affect average costs, as seen in the range of observations, it becomes necessary to cap these claims. This process involves redistributing the burden of claims that exceed a certain threshold, known as the capping threshold. For this purpose, the 95th percentile of the ’average cost’ variable was used as the reference.

Quantiles	
Level	Quantile
100% Max	1099811.20
99%	217650.00
95%	91150.00
90%	61150.00
75% Q3	29230.51
50% Median	6000.00
25% Q1	6000.00
10%	2304.00
5%	1031.43
1%	226.77
0% Min	0.78

Figure 16. Quantiles of the average cost base.

According to Figure 16, average costs equal to or greater than 91 150 were classified as severe claims. This classification enabled us to divide the initial database into two groups: severe claims, comprising 731 observations (see Figure 17), and standard claims, which include the remaining observations.

Analysis Variable				
N	Mean	Std Dev	Minimum	Maximum
731	176101.25	114143.91	91150.00	1099811.20

Figure 17. Simple statistics of the serious claims database (SAS output).

Using "SAS Enterprise Miner", we segmented the standard cost database and recoded the variables according to the identified classes. Table 8 displays the various classes created.

Table 8. The different classes obtained using "SAS Enterprise Miner"

Variables	Classes
Vehicle age (av)	< 5.5 between 5.5 and 13.5 between 13.5 and 20.5 > 20.5
Conductor age (ac)	< 20.5 between 20.5 and 22.5 between 22.5 and 41.5 > 41.5
Circulation zone (zone)	'BeniMellal-Khenifra', 'Guelmim-Oued Noun', 'Marrakech-Safi', 'Casablanca-Settat' , 'Dakhla-Oued Ed Dahab' other regions
Fiscal power (pf)	< 10 other fiscal powers

The tables in Figures 18 and 19 show the number of observations in each class constructed for the previously mentioned variables.

av	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	6706	48.46	6706	48.46
2	3761	27.18	10467	75.65
3	1855	13.41	12322	89.05
4	1515	10.95	13837	100.00

ac	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	34	0.25	34	0.25
2	77	0.56	111	0.80
3	4988	36.05	5099	36.85
4	8738	63.15	13837	100.00

Figure 18. The frequency of observations according to the variable ”av” (left) and according to the variable ”ac” (right).

According to the tables in Figure 18, vehicles were categorized into four groups: the first group contained 6 706 vehicles, while the other groups had 3 761, 1 855, and 1 515 vehicles, respectively. Drivers were also classified into four age groups, with 34, 77, 4 988, and 8 738 individuals in each group, respectively.

zone	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	6235	45.06	6235	45.06
2	7602	54.94	13837	100.00

pf	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	11660	84.27	11660	84.27
2	2177	15.73	13837	100.00

Figure 19. The frequency of observations according to the variable ”zone” (left) and according to the variable ”pf” (right).

The circulation zone was divided into two groups: one with 6 235 observations and the other with 7 602 observations. Fiscal power was also divided into two groups: one with 11 660 observations and the other with 2 177 observations.

4.3. Calculation of the pure premium

This section explores how claim frequency and average costs are influenced by relevant legal frameworks. As highlighted in Remark 3.1, auto insurance models typically address claim frequency and average cost as separate entities. The pure premium is subsequently determined by multiplying these two factors. In practice, claim frequency is commonly modeled using distributions such as the Poisson or Negative Binomial models. For modeling claim costs, the Gamma and log-normal distributions are frequently employed. These continuous distributions are characterized by their definition over the positive real numbers, allowing for a detailed representation of variability in both frequency and cost.

4.3.1. Frequency of claims

After performing the segmentation of rating variables, we undertook the task of identifying the most appropriate distribution to model the frequency of claims, our variable of interest. To evaluate the fit of the proposed models, we utilized the deviance statistic, which measures the discrepancy in log-likelihood between the saturated model (which achieves the maximum fit) and the model being assessed. The deviance provides a comparative metric that helps in determining how well the model captures the underlying data structure.

Assessing model fit is crucial for ensuring reliable statistical inference. A model with a lower deviance relative to the degrees of freedom indicates a better fit to the data, suggesting that it more accurately represents the variability in the frequency of claims. Therefore, when comparing models, it is important to favor those with a lower deviance-to-degrees-of-freedom ratio, as this reflects a more credible and effective model in explaining the observed data.

The deviances of the two distributions are calculated using SAS software.

Criteria For Assessing Goodness Of Fit			
Criterion	DF	Value	Value/DF
Deviance	13E4	780489.9314	6.1923
Scaled Deviance	13E4	780489.9314	6.1923
Pearson Chi-Square	13E4	1547364438.9	12276.5780
Scaled Pearson X2	13E4	1547364438.9	12276.5780
Log Likelihood		-256477.1733	
Full Log Likelihood		-408920.5784	
AIC (smaller is better)		817861.1567	
AICC (smaller is better)		817861.1585	
BIC (smaller is better)		817958.6012	

Criteria For Assessing Goodness Of Fit			
Criterion	DF	Value	Value/DF
Deviance	13E4	65942.2478	0.5232
Scaled Deviance	13E4	65942.2478	0.5232
Pearson Chi-Square	13E4	317039.6994	2.5153
Scaled Pearson X2	13E4	317039.6994	2.5153
Log Likelihood		-47047.9103	
Full Log Likelihood		-47864.3626	
AIC (smaller is better)		95750.7253	
AICC (smaller is better)		95750.7274	
BIC (smaller is better)		95857.9142	

Figure 20. The Poisson distribution deviance (left) and the Negative Binomial distribution deviance (right).

Based on the results from Figure 20, we conducted a thorough assessment of the fit quality for both the Poisson and Negative Binomial distribution models. The key diagnostic metric used was the deviance-to-degrees-of-freedom ratio, which is commonly employed to assess model fit. For the Poisson distribution, we observed that this ratio was significantly higher than that of the Negative Binomial model, indicating a poor fit to the data. Specifically, a higher deviance-to-degrees-of-freedom ratio suggests that the Poisson model is not adequately capturing the overdispersion present in the claim frequency data, where the variance exceeds the mean—a characteristic of real-world insurance claim data. In contrast, the Negative Binomial distribution, which is designed to model overdispersed count data, provided a better fit, as evidenced by its lower deviance-to-degrees-of-freedom ratio. This suggests that the Negative Binomial distribution was better equipped to account for the heterogeneity and variability in claim frequency. As a result, we chose to use the Negative Binomial model exclusively for assessing the significance of the exogenous variables, as it was more appropriate for the structure of the data and provided more reliable estimates for the relationship between predictors and claim frequency.

Analysis Of Maximum Likelihood Parameter Estimates							
Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits	Wald Chi-Square	Pr > ChiSq
Intercept		1	-2.1238	0.0231	-2.1691 -2.0786	8459.01	<.0001
Sex	F	1	0.2077	0.0217	0.1651 0.2502	91.57	<.0001
Sex	M	0	0.0000	0.0000	0.0000 0.0000	.	.
zone	1	1	0.0208	0.0165	-0.0115 0.0531	1.59	0.2071
zone	2	0	0.0000	0.0000	0.0000 0.0000	.	.
Com	G	1	-0.1194	0.0185	-0.1556 -0.0832	41.72	<.0001
Com	D	0	0.0000	0.0000	0.0000 0.0000	.	.
pf	1	1	0.1965	0.0231	0.1513 0.2417	72.59	<.0001
pf	2	0	0.0000	0.0000	0.0000 0.0000	.	.
ac	1	1	0.1473	0.0279	0.0926 0.2021	27.82	<.0001
ac	2	1	-0.0082	0.0170	-0.0415 0.0250	0.23	0.6282
ac	3	0	0.0000	0.0000	0.0000 0.0000	.	.
av	1	1	1.1301	0.0248	1.0814 1.1788	2068.49	<.0001
av	2	1	0.6968	0.0229	0.6520 0.7416	929.68	<.0001
av	3	1	0.2933	0.0288	0.2368 0.3499	103.39	<.0001
av	4	0	0.0000	0.0000	0.0000 0.0000	.	.
Dispersion		0	0.0000	0.0000	0.0000 0.0000		

Figure 21. Results of WALD significance test for Negative Binomial (SAS output).

We used the Wald significance test to estimate the parameters of the Negative Binomial distribution model, and we applied the GLM in SAS using the *Genmod* procedure, which produced the following results.

As shown in Figure 21, the majority of the included variables significantly contribute to explaining claim frequency, with p-values below the 5% threshold. However, two variables—"zone" and "ac2"—were found to be statistically insignificant, indicating that their inclusion did not meaningfully improve the explanatory power of the model. These variables were therefore excluded from the final model to enhance its efficiency and interpretability.

The rationale for initially including these variables lies in their theoretical relevance and potential impact on claim frequency. For example, the "zone" variable was included to capture potential geographic effects, given that urban and rural areas may exhibit different risk profiles. Similarly, the "ac2" variable was hypothesized to contribute based on its correlation with other key factors. However, the empirical results revealed that these variables did not significantly affect claim frequency in our dataset.

After removing "zone" and "ac2," we re-estimated the model, and the updated results are presented in Figure 22. This revised model demonstrated improved statistical performance, with no loss in explanatory power, thereby supporting the decision to exclude these variables.

Analysis Of Maximum Likelihood Parameter Estimates

Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	−2.1160	0.0222	−2.1596	−2.0724	9062.80	<.0001
Sex	F	1	0.2076	0.0217	0.1650	0.2501	91.48	<.0001
Sex	M	0	0.0000	0.0000	0.0000	0.0000	.	.
Com	G	1	−0.1193	0.0185	−0.1556	−0.0831	41.67	<.0001
Com	D	0	0.0000	0.0000	0.0000	0.0000	.	.
pf	1	1	0.1964	0.0231	0.1512	0.2416	72.49	<.0001
pf	2	0	0.0000	0.0000	0.0000	0.0000	.	.
ac	1	1	0.1475	0.0279	0.0927	0.2022	27.87	<.0001
av	1	1	1.1301	0.0248	1.0814	1.1788	2068.45	<.0001
av	2	1	0.6966	0.0229	0.6519	0.7414	929.30	<.0001
av	3	1	0.2933	0.0288	0.2368	0.3499	103.38	<.0001
av	4	0	0.0000	0.0000	0.0000	0.0000	.	.
Dispersion		1	0.0000	0.0002	0.0000	3.56E127		

Figure 22. Results of the Wald significance test for NB without the variables "zone" and "ac2" (SAS output).

4.3.2. The average cost of claims

In this section, we aim to identify the most appropriate distribution for accurately modeling the average cost of claims. Selecting an appropriate statistical distribution is crucial for ensuring that the model provides reliable predictions and captures the underlying patterns in the data. To achieve this, we employ a rigorous assessment of model fit, leveraging the deviance as a key evaluation metric. This approach enables us to compare the performance of different candidate distributions in terms of their ability to represent the observed data.

As in the preceding section, the primary criterion for evaluating the quality of the fit is the ratio of deviance to degrees of freedom. This metric allows for a standardized comparison of models, taking into account the complexity of each distribution. A lower deviance-to-degrees-of-freedom ratio indicates a better fit, as it reflects a closer alignment between the predicted and observed values with minimal overfitting.

For this analysis, we focus on two widely used distributions in the modeling of claim costs: the Gamma distribution and the log-normal distribution. Both distributions are well-suited for modeling positive continuous data, making them natural candidates for representing inherently non-negative claim costs.

Criteria For Assessing Goodness Of Fit			
Criterion	DF	Value	Value/DF
Deviance	14E3	17847.0491	1.2910
Scaled Deviance	14E3	16158.3849	1.1689
Pearson Chi-Square	14E3	19342.1397	1.3992
Scaled Pearson X2	14E3	17512.0120	1.2668
Log Likelihood		-148042.5383	
Full Log Likelihood		-148042.5383	
AIC (smaller is better)		296099.0766	
AICC (smaller is better)		296099.0847	
BIC (smaller is better)		296151.8187	

Criteria For Assessing Goodness Of Fit			
Criterion	DF	Value	Value/DF
Deviance	14E3	21124.2585	1.5284
Scaled Deviance	14E3	13830.0000	1.0007
Pearson Chi-Square	14E3	21124.2585	1.5284
Scaled Pearson X2	14E3	13830.0000	1.0007
Log Likelihood		-22552.9889	
Full Log Likelihood		-22552.9889	
AIC (smaller is better)		45125.9779	
AICC (smaller is better)		45125.9938	
BIC (smaller is better)		45201.3238	

Figure 23. The Gamma distribution deviance (left) and the log-normal distribution deviance (right).

Based on the figures above (Figure 23), the goodness-of-fit test confirms that the model with a Gamma distribution fits our data better than the model with a log-normal distribution. Therefore, we will continue our analysis by focusing exclusively on the Gamma distribution model to assess the significance of the exogenous variables, followed by residual tests.

As in the ”frequency” section, we used the WALD significance test to estimate the parameters of the Gamma distribution model. Similarly, we eliminated variables with p-values greater than 5%, and we obtained the following results.

Analysis Of Maximum Likelihood Parameter Estimates							
Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits	Wald Chi-Square	Pr > ChiSq
Intercept		1	10.0232	0.0279	9.9684 10.0779	128740	<.0001
Sex	F	1	-0.1429	0.0233	-0.1886 -0.0971	37.46	<.0001
Sex	M	0	0.0000	0.0000	0.0000 0.0000	.	.
av	1	1	-0.4384	0.0303	-0.4978 -0.3790	209.20	<.0001
av	2	1	-0.3149	0.0322	-0.3779 -0.2518	95.88	<.0001
ac	1	1	0.0416	0.0186	0.0051 0.0780	4.99	0.0254
ac	2	0	0.0000	0.0000	0.0000 0.0000	.	.
Scale		1	0.9054	0.0095	0.8869 0.9242		

Figure 24. Results of WALD significance test for Gamma (SAS output).

To ensure the reliability and robustness of the model, testing the normality of the residuals plays a pivotal role. Residual analysis is a critical diagnostic tool that helps identify discrepancies between predicted values and observed data, providing insights into the model's adequacy and highlighting potential areas of misfit. By examining residual patterns, it becomes possible to assess whether the underlying assumptions of the model hold true, thereby enhancing confidence in the results.

In this study, the analysis was performed using the GLM framework in SAS, specifically leveraging the ”*proc genmod*” procedure. This procedure provides a comprehensive platform for fitting GLMs, offering robust tools for evaluating residuals, estimating parameters, and testing hypotheses. By employing this approach, we were able to systematically examine the residuals and assess whether they align with the assumptions of the chosen model distribution.

The results of this analysis are summarized in Figure 25. This figure illustrates the distribution of the deviance residuals, providing a visual representation of their alignment with normality.

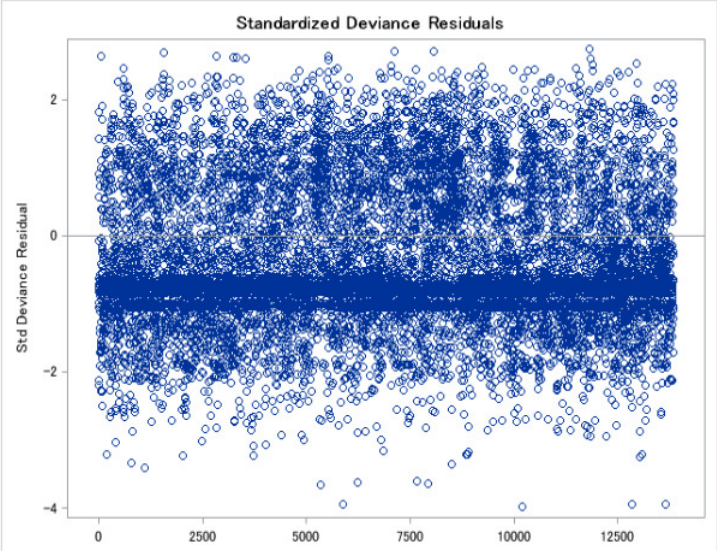

Figure 25. Standardized deviance residuals (SAS output).

The residuals are fairly randomly distributed around 0, suggesting that the linearity assumption underlying the model is acceptable. Additionally, the residuals form a horizontal band around 0, indicating that the errors ε_i have constant variance. In practice, for the model to be validated, the residuals should fall within the interval $[-2.5, 2.5]$, with no more than 5% of the residuals outside this range. In our case, this assumption is confirmed.

5. Discussion

Our study demonstrated that the Negative Binomial distribution and the Gamma distribution are the most appropriate for modeling frequency and average cost, respectively, using the GLM. The premium for each segment was calculated as the product of the average cost and frequency. The final models are therefore presented as follows.

$frequency = exhibition * exp(-2.1160 + 0.2076 * (Sex\ in\ ('F')) - 0.1193 * (Com\ in\ ('G')) + 0.1964 * (pf\ in\ ('10-14')) + 0.1475 * (ac < 30.5) + 1.1301 * (av < 2.5) + 0.6966 * (av >= 2.5\ and\ av < 10.5) + 0.2933 * (av >= 10.5\ and\ av < 16.5));$

$averagecost = exp(10.02332 - 0.1429 * (sex\ in\ ('F')) - 0.4384 * (av < 5.5) - 0.3149 * (av >= 5.5\ and\ av < 13.5) + 0.0454 * (ac < 41.5));$

The pure premium is given by: PP = Frequency $\times$ Average cost + Average cost of serious loss $\times\Pi$, with:

- Average cost of serious loss = 176101.25 (see Figure 17).
- Π=number of insured persons who had a serious loss / number of insured persons = 731/126082=0.00579781.

Hence, PP = Frequency $\times$ Average cost + 1021.

Example Output (Simulation)

To show an example of the output, assume the following values: "exhibition = 1", "sex = F", "com = G", "pf = 12", "ac = 25", and "av = 3".

1. Frequency calculation:
   $frequency = 1 \times exp(-2.1160 + 0.2076 \times 1 - 0.1193 \times 1 + 0.1964 \times 1 + 0.1475 \times 1 + 1.1301 \times 0 + 0.6966 \times 1 + 0.2933 \times 0) = exp(-0.8972) \backsimeq 0.4078.$

2. Average cost calculation:
$averagecost = exp(10.02332 - 0.1429 \times 1 - 0.4384 \times 0 - 0.3149 \times 0 + 0.0454 \times 1) = exp(9.92582) \backsimeq 20501.9036$.

Hence, PP $= 0.4078 \times 20501.9036 + 1021 = 9381.6762$.

This study provides a robust application of GLMs for insurance pricing; however, it is not without limitations. A notable constraint is the dataset employed, which, despite being representative, did not include key predictors such as driving history, vehicle type, and policy characteristics. These variables could potentially enhance the model's accuracy and generalizability. Addressing these limitations in future research would require access to more comprehensive and high-quality data.

Implementing GLMs in the insurance industry also presents practical challenges. Issues such as data quality, computational demands, and regulatory constraints can impact the model's effectiveness. Ensuring data consistency, investing in computational resources, and adhering to regulatory requirements are essential steps to enhance the reliability and interpretability of GLMs for stakeholders. Furthermore, model validation across diverse contexts and regions remains critical for assessing their robustness and applicability in various insurance markets.

We acknowledge that validating the models with datasets from other countries or regions would significantly enhance the generalizability and robustness of our findings. While this aspect was not included in the current study due to the project's initial scope, it remains a valuable avenue for future research. Such validation efforts could provide insights into the adaptability of GLMs across various insurance markets and ensure their reliability in diverse regulatory and demographic contexts.

Future research should also expand the scope of this study by integrating additional explanatory variables, including traffic conditions, driving behavior, and macroeconomic indicators, to better capture the factors influencing claim frequency and severity. The exploration of dynamic pricing strategies, incorporating real-time data such as telematics and customer-centric metrics, would align with emerging industry trends and provide a more adaptive approach to pricing optimization.

Moreover, a comparative analysis of GLMs with other advanced modeling techniques, such as machine learning algorithms and generalized additive models, could offer insights into their relative performance. Such an analysis would strengthen the case for adopting GLMs or reveal alternative approaches that might outperform them under specific conditions. This comparison, alongside further exploration of regulatory impacts and industry-specific constraints, could yield innovative solutions for pricing optimization and risk management in the insurance sector.

6. Conclusion

This study addressed the critical issue of pricing in auto insurance, with a focus on determining an accurate premium for policyholders to secure coverage against risk. Given the dynamic nature of the insurance sector, marked by evolving market competition, regulatory changes, and economic conditions, it is crucial to regularly and precisely review premium calculations. In this paper, we applied GLM to our auto insurance portfolio to predict claim frequencies and severities by modeling the relationship between various risk factors and the cost of claims. The implementation process involved several key steps: data processing and cleaning of production and claims databases, conducting a descriptive analysis of the utilized data, segmenting pricing variables, and selecting appropriate distributions for the two primary variables: claim frequency and average claim cost. Specifically, the Negative Binomial distribution was chosen to model claim frequency, while the Gamma distribution was selected for average claim cost. Each model underwent rigorous statistical testing, including the Wald and deviance tests. The validation of these models confirmed the accuracy of the estimates. In conclusion, the experimental findings underscored the significance of GLMs in actuarial science, particularly in refining insurance pricing strategies.

REFERENCES

1. J. A. Nelder, and R. W. Wedderburn, *Generalized linear models*, Journal of the Royal Statistical Society Series A: Statistics in Society, vol. 135, no. 3, pp. 370–384, 1972.
2. J. M. Wooldridge, *Econometric analysis of cross section and panel data*, MIT press, 2010.
3. A. C. Cameron, and P. K. Trivedi, *Regression analysis of count data*, Cambridge university press, no. 13, 2013.
4. M. Verbeek, *A guide to modern econometrics*, John Wiley & Sons, 2017.
5. D. Collett, *Modelling binary data*, CRC press, 2002.
6. D. W. Hosmer Jr, S. Lemeshow, and R. X. Sturdivant, *Applied logistic regression*, John Wiley & Sons, 2013.
7. A. J. Dobson, and A. G. Barnett, *An introduction to generalized linear models*, Chapman and Hall/CRC, 2018.
8. J. Scott Long, *Regression models for categorical and limited dependent variables*, Advanced quantitative techniques in the social sciences vol. 7, 1997.
9. A. Agresti, *Categorical data analysis*, John Wiley & Sons, vol. 792, 2012.
10. P. McCullagh, *Generalized linear models*, Routledge, 2019.
11. R. H. Myers, D. C. Montgomery, G. G. Vining, and T. J. Robinson, *Generalized linear models: with applications in engineering and the sciences*, John Wiley & Sons, 2012.
12. D. C. Montgomery, and G. C. Runger, *Applied statistics and probability for engineers*, John wiley & sons, 2020.
13. H. D. Skipper, *Risk management and insurance: perspectives in a global economy*, John Wiley & Sons, 2008.
14. E. W. Frees, *Regression modeling with actuarial and financial applications*, Cambridge University Press, 2009.
15. G. Plantin, and J. C. Rochet, *When insurers go bust: an economic analysis of the role and design of prudential regulation*, Princeton University Press, 2016.
16. P. Zweifel, R. Eisen, and D. L. Eckles, *Insurance markets and asymmetric information*, Insurance Economics, pp. 315–381, 2021.
17. J. Burling, and K. Lazarus, *Research handbook on international insurance law and regulation*, Edward Elgar Publishing, 2023.
18. Supervisory Authority for Insurance and Social Welfare (ACAPS). https://www.acaps.ma/en.
19. M. Slime, M. El Kamli, and A. Ould Khal, *Dependence Modeling in Non-Life Insurance: Copula Functions and Capital Adequacy-A Case Study of AXA Insurance*, IAENG International Journal of Applied Mathematics, vol. 54 no. 4, 2024.
20. S. E. Harrington, and G. Niehaus, *Risk management and insurance*, 1999.
21. G. E. Rejda, *Principles of risk management and insurance*, Pearson Education India, 2005.
22. A. J. McNeil, R. Frey, and P. Embrechts, *Quantitative risk management: concepts, techniques and tools-revised edition*, Princeton university press, 2015.
23. R. Sachs, *Risk and Uncertainty in the Insurance Industry*, Psychological Perspectives on Risk and Risk Analysis. Springer, 2018.
24. S. Asmussen, and M. Steffensen, *Risk and insurance*, Springer International Publishing, 2020.
25. P. De Jong, and G. Z. Heller, *Generalized linear models for insurance data*, Cambridge University Press, 2008.
26. E. W. Frees, R. A. Derrig, and G. Meyers, *Predictive modeling applications in actuarial science*, Cambridge University Press, vol. 1, 2014.
27. V. I. Rotar, *Actuarial models: the mathematics of insurance*, CRC Press, 2014.
28. S. Kafková, and L. Křivánková, *Generalized linear models in vehicle insurance*, Acta Universitatis Agriculturae et Silviculturae Mendelianae Brunensis, vol. 62, no. 2, pp. 383–388, 2014.
29. M. David, *Auto insurance premium calculation using generalized linear models*, Procedia Economics and Finance, vol. 20, pp. 147–156, 2015.
30. S. Xie, and A. T. Lawniczak, *Estimating major risk factor relativities in rate filings using generalized linear models*, International Journal of Financial Studies, vol. 6, no. 4, pp. 84, 2018.
31. E. Šoltés, S. Zelinová, and M. Bilíková, *General linear model: an effective tool for analysis of claim severity in motor third party liability insurance*, STATISTICS, vol. 13, 2019.
32. E. Seyam, and H. Elsalmouny, *Proposed models for comprehensive automobile insurance ratemaking in Egypt with parametric and semi-parametric regression: a case study*, Journal of Statistics Applications & Probability, vol. 11, no. 1, pp. 41–55, 2022.
33. R. OKTAVIA, R. ZUHRA, H. HAFNANI, N. NURMAULIDAR, and I. SYAHRINI, *Application of Poisson and negative binomials models to estimate the frequency of insurance claims*, Jurnal Natural, vol. 23, no. 1, pp. 21–27, 2023.
34. E. Brati, *Application of GLM and GAMLSS Models in Predictive Analysis of Motor Bodily Injury Claims*, International Conference on Business and Technology. Cham: Springer Nature Switzerland, pp. 365–375, 2024.
35. Moroccan Federation of Insurance and Reinsurance Companies (FMSAR). https://fma.org.ma/en.