Lisbon School of Economics & Management
Universidade de Lisboa

# MASTER
## MATHEMATICAL FINANCE

# MASTER´S FINAL WORK
## INTERNSHIP REPORT

## HARVESTING RISK: APPLYING GLMS TO AGRICULTURAL INSURANCE PRICING

BERNARDO DIAS BENTO

**SUPERVISION:**
PROF. ALEXANDRA MOURA
ANTÓNIO LOBO

JUNE 2025

# Acknowledgements

I begin writing this chapter with mixed emotions—pride, exhaustion, and happiness bubbling to the surface as I reflect on this significant milestone.

I want to thank my family for always believing in me.
To my mother, Graça, and my father, Rui, for all the support and dedication — one day I hope to repay you for all you´ve to me. To my grandmother, Arciolinda, all my uncles and cousins, to my mother-in-law Isabel, my brother-in-law Pedro, and my nephews and nieces, Simão, Maria, and Constança.
A very special thank you to my sister, Andreia, who has always been a role model—someone I look up to and aspire to become.

A heartfelt thank you to Catarina, my girlfriend, best friend, and life partner. A simple "thank you" will never be enough for everything you've done—and continue to do—for me. You've always been my rock and, above all, you've given me the strength to keep going and never give up.

Thank you to Gordicho, Rita, Alegria, Jéssica, Marcos, Joana, Coelho, Bea, Banha, Ângelo, Raquel, and Artur, who stood by me through the happiest and toughest moments. I'm deeply grateful for your friendship and the care you've always shown me.

To João and Miguel, who've put up with me and supported me since our undergraduate days—thank you for your friendship and for making these past five years unforgettable.

To all my colleagues at Atlas and Secose whom I had the pleasure of meeting during these last four months—thank you for welcoming me with a smile and a helping hand, always ready to assist in whatever I needed.

My deepest thanks to Professor Alexandra Moura, not only for supervising my dissertation, but also for all the ideas, clarifications, revisions, and, above all, for the trust she placed in me.

Finally, a big thank you to Frederico and António for the opportunity to be part of such an incredible project and company. Words will never be enough to express my gratitude for the trust you placed in me and for giving me the chance to grow through this work.

*I'm very fortunate. Thank you all...*

# Resumo

O seguro agrícola desempenha um papel fundamental na mitigação dos riscos financeiros enfrentados pelos agricultores devido a eventos climáticos adversos. Um dos fatores essenciais de qualquer produto de seguro é a estimativa do prémio puro, que reflete o custo esperado dos sinistros. Este trabalho tem como objetivo avaliar e melhorar o processo de estimativa do prémio puro utilizado na Atlas MGA, companhia de seguros portuguesa especializada em seguros agrícolas.

Atualmente, a empresa utiliza uma abordagem baseada em dados históricos (do IFAP) e num ratio chave. Neste trabalho, é proposta a aplicação do GLM, uma alternativa mais flexível e mais robusta.

O prémio puro é calculado através de duas componentes: a frequência esperada de sinistros e a severidade esperada de cada um deles. Os modelos foram implementados utilizando dados reais de seguros agrícolas fornecidos pelo IFAP e pela companhia, e os resultados foram comparados com o método atualmente utilizado pela empresa. Os resultados demonstram que os GLMs fornecem estimativas de prémio mais precisas e consistentes, e sugerem que o método tradicional pode subestimar sistematicamente o prémio puro em alguns casos.

Este estudo evidencia o potencial dos GLMs para aprimorar modelos atuariais de precificação em seguros agrícolas e fornece uma base para futuras melhorias, incluindo a incorporação de variáveis adicionais e o uso de técnicas preditivas mais avançadas, tais como as redes neurais ou o machine learning.

**Palavras-chave:** Prémio Puro, Seguro Agrícola, Modelos Lineares Generalizados (GLM), Severidade, Frequência, Precificação Atuarial.

# Abstract

Agricultural insurance plays a fundamental role in mitigating the financial risks faced by farmers due to adverse climatic events. One of the key elements of any insurance product is the estimation of the pure premium, which reflects the expected cost of claims. This study aims to evaluate and improve the pure premium estimation process currently used by Atlas MGA, a portuguese insurance company specialized in agriculture insurance.

At present, the company adopts an approach to premium pricing that is based on historical data (from IFAP) on a key ratio. In this work, we propose the application of Generalized Linear Models (GLMs) as a more flexible and more robust alternative.

The pure premium is calculated through two components: the expected frequency of claims and the expected severity each claims. The models were implemented using real agricultural insurance data provided by IFAP and the company, and the results were compared with the method currently used by the company. The findings show that GLMs produce more accurate and consistent premium estimates, and suggest that the traditional method may systematically underestimate the pure premium in certain cases.

This study highlights the potential of GLMs to enhance actuarial pricing models in agricultural insurance and provides a foundation for future improvements, including the incorporation of additional variables and the use of more advanced predictive techniques, such as neural networks and machine learning.

**Keywords:** Pure Premium, Agricultural Insurance, Generalized Linear Models (GLM), Severity, Frequency, Actuarial Pricing.

# Disclaimer

This master internship report was developed with strict adherence to the academic integrity policies and guidelines set forth by ISEG, Universidade de Lisboa. The work presented herein is the result of my own research, analysis, and writing, unless otherwise cited. In the interest of transparency, I provide the following disclosure regarding the use of artificial intelligence (AI) tools in the creation of this thesis/internship report/project:

I disclose that AI tools were employed during the development of this thesis for translation, literature review assistance, and support with LaTeX software. However, all final writing, synthesis, and critical analysis are my own work.

Nonetheless, I have ensured that the use of AI tools did not compromise the originality and integrity of my work. All sources of information, whether traditional or AI-assisted, have been appropriately cited in accordance with academic standards. The ethical use of AI in research and writing has been a guiding principle throughout the preparation of this thesis.

# Table Of Contents

# List of Figures

# List of Tables

# List of Abbreviations and Acronyms

| | |
|---|---|
| **GLM** | Generalized Linear Models |
| **AIC** | Akaike Information Criterion |
| **FLE** | Fire, Lightning, Explosion |
| **FST** | Frost |
| **HAI** | Hail |
| **OTH** | Other |
| **PR** | Persistent Rain |
| **SNW** | Snow |
| **SUN** | Sunscald |
| **TOR** | Tornado |
| **WTR** | Waterspout |
| **IFAP** | Instituto de Financiamento da Agricultura e Pescas,I.P |

# 1 Introduction

Agriculture plays a fundamental role in our society. However, agricultural producers are constantly exposed to climatic and natural risks, which pose a threat to their crops. In this context, agricultural insurance emerges as an important tool to protect farmers against these risks, providing financial security and helping to mitigate potential losses. This study is entirely focused on the pricing and estimation of the pure premium of agriculture liabilities. The pure premium is the amount required to cover the expected cost of claims.

Atlas MGA, a company specialized in agricultural insurance, currently uses a simplified methodology to calculate the pure premium. This approach is based on a key ratio and historical data from IFAP on claims and insured capital. However, we believe this methodology can be improved through more sophisticated statistical techniques, such as Generalized Linear Models (GLMs), which have proven effective in various areas of non-life insurance pricing. The main question to be explored is whether adopting a GLM-based model can provide a more accurate and robust estimate of the pure premium when compared to the company's current method.

The main objective of this work is to develop and evaluate a GLM model to estimate the pure premium of agricultural insurance and compare it with the model currently used by Atlas MGA. To achieve this goal, two key aspects inherent to actuarial pricing will be addressed: the modeling of claim frequency and claim severity (i.e., average cost of claims). The use of a GLM provides a more flexible and efficient framework, capable of handling various types of distributions while capturing new market trends and climate-related events.

To achieve the proposed objectives, this work is structured as follows: in **Chapter 2**, the dataset used will be presented, along with its preprocessing steps. In **Chapter 3**, the methodology will be detailed, providing the theoretical background for the approach used. In **Chapter 4**, the results of the GLM application will be presented, including a comparison with the company's current pricing method. Finally, **Chapter 5** presents the conclusions of the study and suggestions for future research.

# 2 Context and Motivation

Insurance is a concept deeply embedded in society. It is a contract between an insurer and a policyholder, where the former agrees to cover significant losses of the latter's assets in exchange for the payment of a premium. While insurance types such as automobile, health, or home are common in everyday life, this dissertation focuses on a specific category: agricultural insurance.

As the name suggests, crop insurance is a type of agricultural insurance designed to protect farmers from significant losses in their harvests, primarily due to natural events such as rain, hail, frost, among others.

This is the main area of expertise of the company where this project was developed: Atlas MGA. Atlas is a Managing General Agent (MGA) specialized in agricultural insurance, and one of the leading players in the Iberian Peninsula. The policies offered by Atlas cover a variety of risks related to natural events. Each crop is associated with a specific insurance product, with a set of mandatory coverages, to which additional coverages can be added. For instance, vineyard crop insurance includes as mandatory coverages: Fire, Lightning, Explosion, Snow, Hail, Persistent Rain, Tornado, and Waterspout. Additional coverages can include Frost and Sunscald. For this reason, the premium calculation must be performed by coverage, the final pure premium being obtained as the sum of the values of the selected coverages.

The objective of this work was to contribute to improving the pricing model for the company's pure premium estimation. Currently, the pure premium is calculated using a relatively simple and direct approach:

$$PureTax = \frac{\sum \text{Claims}}{\sum \text{Insured Capital}} \tag{1}$$

where $PureTax$ is the ratio that reflects the average amount paid in claims by the insurer per monetary unit of insured capital, $\sum$ Claims represents the total amount of claims over the observed years, and $\sum$ Insured Capital corresponds to the total insured amounts for those same years. In agricultural insurance, the insured capital is commonly calculated as:

$$\text{Insured Capital} = \text{Area} \times \text{Average Productivity} \times \text{Product Price}$$

The pure premium is estimated, within the company, using historical claims data for a given municipality and crop. This is done by calculating the total amount of claims and the total insured capital over the insured years. The resulting value from equation (1) is then multiplied by the insured capital of the new client to obtain the estimated pure premium.

In this study, we consider one of the most widely used methodologies in insurance pricing today: Generalized Linear Models (GLMs). Introduced in 1972 by John Nelder and Robert Wedderburn, GLMs represented a significant contribution to the field of statistics, expanding the range of regression models beyond the classical linear (normal), logistic, and Poisson regressions.

The main objective of this study is to develop a more robust and accurate method for estimating the pure premium compared to the approach currently used by the company.

# 3 Dataset

This chapter presents the data used in this study, detailing its source, the preparation procedures, and the justification behind the selection of the distributions employed in the Generalized Linear Model (GLM). These steps are fundamental to ensure that the modeling process is based on reliable, well-structured, and appropriately treated data.

## 3.1 Data Source

All data related to agricultural insurance policies and claims were provided by IFAP (Instituto de Financiamento da Agricultura e Pescas). IFAP is the Portuguese public entity responsible for managing and distributing funds allocated to agriculture, rural development, and fishing.

IFAP, as a public entity, supports the uptake of agricultural insurance by subsidizing part of the premium cost. In most cases, IFAP covers 60% of the commercial premium, while the farmer is responsible for the remaining 40%, in addition to taxes and fees (11% of the commercial premium). IFAP subsequently reimburses the subsidized portion directly to the insurance company.
This public co-financing of agricultural insurance is justified by the need to protect farming activities against unpredictable climatic risks. It aims to promote income stability for producers, ensure continuity in agricultural production, and strengthen the resilience of the national agricultural sector.

The dataset consists of 21,208 records from IFAP, publicly available under request, organized by year, crop type, and municipality. This means that the data do not correspond to individual policyholders but rather to the aggregation of all insured parties within a municipality for a given crop. Each record provides the total insured capital for that municipality and crop, along with the corresponding loss amount in case of a claim. Additionally, a supplementary table is provided detailing the cause of the loss and how much each type of loss contributed to the total claim amount. Tables 1 and 2 present examples of the main dataset and the associated loss cause table, respectively.

| Key | Year | Crop | Municipality | Sum Insured | Losses |
|---|---|---|---|---|---|
| 1996_Ameixa_BELMONTE_D | 1996 | Ameixa | BELMONTE | 1 800 000,00 € | 0,00 € |
| 1996_Ameixa_CAMPO MAIOR_C | 1996 | Ameixa | CAMPO MAIOR | 18 710 000,00 € | 1 608 000,00 € |
| 1996_AVEIA_EVORA_C | 1996 | AVEIA | EVORA | 161 049 790,00 € | 413 126,00 € |

Table 1: Example of the general dataset provided by IFAP

| Key | Year | Crop | Municipality | Cause | Loss |
|---|---|---|---|---|---|
| 1996_Ameixa_CAMPO MAIOR_C | 1996 | Ameixa | CAMPO MAIOR | Hail | 1 608 000,00 € |
| 1996_AVEIA_EVORA_C | 1996 | AVEIA | EVORA | Hail | 211 848,00 € |
| 1996_AVEIA_EVORA_C | 1996 | AVEIA | EVORA | Fire, Lightning, explosion | 201 278,00 € |

Table 2: Example of the table detailing loss causes

As previously mentioned, the primary objective of this work is to estimate the pure premium of an insurance policy for a given crop in a specific municipality. After appropriate preprocessing, these two datasets will constitute the main source of data used for the construction of the GLM.

Table 3 presents the description of the variables used in the study's dataset.

| Variable | Description | Additional Information |
|---|---|---|
| YEAR | Year | From year 2002 to 2021 |
| CROP | Crop Type | 112 Different Types |
| MUNICIPALITY | Municipality | 268 Different Types |
| SUMINS | Sum Insured | Range between 4 and 23477256.94 |
| LOSS | Loss Amount | Range between 0 and 5127950,56 |
| CAUSE | Cause of Claim | 9 Different Types + Type "None" |
| FLE | Fire, Lightning, Explosion | Table6 |
| FST | Frost | Table6 |
| HAI | Hail | Table6 |
| OTH | Other | Table6 |
| PR | Persistent Rain | Table6 |
| SNW | Snow | Table6 |
| SUN | Sunscald | Table6 |
| TOR | Tornado | Table6 |
| WTR | Waterspout | Table6 |

Table 3: Description of the Variables in the Dataset

## 3.2 Data Processing

Based on the datasets provided by IFAP, a preprocessing procedure was carried out to prepare the data for analysis. Two distinct tables were constructed: one for losses and another for the frequency of each type of claim.

- **Loss Table**

  This table closely resembles the general dataset provided by IFAP, but it includes only data from 2002 onwards, due to the currency transition from the Portuguese Escudo to the Euro in January 2002. Given the uncertainty regarding the exact point when insured and loss amounts began to be recorded in euros, we chose to consider only data from that year forward, resulting in a total of 17,842 records, covering the period from 2002 to 2021.

  The resulting dataset is organized into five columns: YEAR (the year of the crop), CROP (the type of crop), MUNICIPALITY (the municipality), SUMINS (the total insured capital for that municipality and crop), and LOSS (the corresponding loss amount in the event of a claim). Table 4 presents an excerpt of this table.

| KEY | YEAR | CROP | MUNICIPALITY | SUMINS | LOSS |
|---|---|---|---|---|---|
| 2002_ABACATE_LOULE_A | 2002 | ABACATE | LOULE | 14 875,00 € | 0,00 € |
| 2002_ABÓBORA_VILA FRANCA DE XIRA_B | 2002 | ABÓBORA | VILA FRANCA DE XIRA | 12 000,00 € | 0,00 € |
| 2002_AMEIXA_ALANDROAL_C | 2002 | AMEIXA | ALANDROAL | 13 500,00 € | 0,00 € |
| 2002_AMEIXA_ALCOBACA_B | 2002 | AMEIXA | ALCOBACA | 23 217,92 € | 4 800,68 € |
| 2002_AMEIXA_ARMAMAR_D | 2002 | AMEIXA | ARMAMAR | 28 210,00 € | 1 107,20 € |
| 2002_AMEIXA_BELMONTE_D | 2002 | AMEIXA | BELMONTE | 18 505,00 € | 3 463,20 € |
| 2002_AMEIXA_BOMBARRAL_B | 2002 | AMEIXA | BOMBARRAL | 239 575,00 € | 3 051,34 € |
| 2002_AMEIXA_BORBA_C | 2002 | AMEIXA | BORBA | 195 210,00 € | 6 488,88 € |
| 2002_AMEIXA_BRAGA_D | 2002 | AMEIXA | BRAGA | 4 000,00 € | 1 853,13 € |

Table 4: Excerpt of the Loss Table

- **Claim Type Table**

  This table was specifically constructed to capture the frequency of each type of claim. There are eight primary types of claims: Fire, Lightning, Explosion (FLE); Frost (FST); Hail (HAI); Persistent Rain (PR); Snow (SNW); Sunscald (SUN); Tornado (TOR); and Waterspout (WTR). Any remaining types were grouped under the category OTHER (OTH).

  The table was constructed in a binary format: for each record, additional columns corresponding to each claim type were added, containing a value of 1 if that type of claim occurred and 0 otherwise. For instance, if the claim was due to frost, the column labeled FST would contain a 1, while all other columns would contain 0. In cases where no claim occurred, all columns are set to 0. Table 5 shows an excerpt of this table, corresponding to the same rows from the loss table shown previously for comparison purposes.

| KEY | YEAR | CULTURE | MUNICIPALITY | CAUSE | FLE | FST | HAI | OTH | PR | SNW | SUN | TOR | WTR |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2002_ABACATE_LOULE | 2002 | ABACATE | LOULE | None | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2002_ABÓBORA_VILA FRANCA DE X | 2002 | ABÓBORA | VILA FRANCA DE XIRA | None | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2002_AMEIXA_ALANDROAL | 2002 | AMEIXA | ALANDROAL | None | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2002_AMEIXA_ALCOBACA | 2002 | AMEIXA | ALCOBACA | Frost | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2002_AMEIXA_ALCOBACA | 2002 | AMEIXA | ALCOBACA | Hail | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2002_AMEIXA_ARMAMAR | 2002 | AMEIXA | ARMAMAR | Frost | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2002_AMEIXA_BELMONTE | 2002 | AMEIXA | BELMONTE | Frost | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2002_AMEIXA_BOMBARRAL | 2002 | AMEIXA | BOMBARRAL | Tornado | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| 2002_AMEIXA_BOMBARRAL | 2002 | AMEIXA | BOMBARRAL | Hail | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2002_AMEIXA_BORBA | 2002 | AMEIXA | BORBA | Hail | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2002_AMEIXA_BRAGA | 2002 | AMEIXA | BRAGA | Frost | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

Table 5: Excerpt of the Claim Type Table

As can be seen, some municipalities experienced multiple claims within a single year. For example, Alcobaça recorded occurrences of both frost (FST) and hail (HAI) in the same year.

## 3.3 Selection of Response and Explanatory Variables

The objective of this study is to employ a Generalized Linear Model (GLM) to estimate both the frequency and severity of claims, considering the region (municipality) and crop type. Consequently, these factors are expected to serve as the primary explanatory variables of the model, alongside the insured capital.

Given the aim of estimating the pure premium for each coverage unit, the response variable for the frequency model will be the occurrence or non-occurrence of each type of claim, resulting in a binary structure. For this reason, the Bernoulli($p$) distribution was selected as the response distribution, with municipality and crop acting as explanatory variables. Insured capital was excluded from the frequency model, as it does not directly influence the likelihood of a claim occurring. Table 6 shows some summary statistics of the variables regarding the claim type.

| Variable | 0's | 1's | Mean | Variance |
|---|---|---|---|---|
| FLE | 21100 | 774 | 0.0354 | 0.0341 |
| FST | 18370 | 3504 | 0.1602 | 0.1345 |
| HAI | 18374 | 3500 | 0.1600 | 0.1344 |
| OTH | 21813 | 61 | 0.0028 | 0.0028 |
| PR | 21841 | 33 | 0.0015 | 0.0015 |
| SNW | 21860 | 14 | 0.0006 | 0.0006 |
| SUN | 21686 | 188 | 0.0086 | 0.0085 |
| TOR | 21629 | 245 | 0.0112 | 0.0111 |
| WTR | 21448 | 426 | 0.0195 | 0.0191 |

Table 6: Descriptive Statistics of Claim Type variables

For the severity model, however, the explanatory variables will include municipality, crop, and insured capital. Since the response variable represents positive continuous values (claim amounts), and following established actuarial practice in GLMs for insurance pricing, the Gamma distribution was chosen as the response distribution. Table 7 shows some summary statistics of the loss variable.

| Variable | Mean | Variance | Min | Max | Skewness | Kurtosis |
|---|---|---|---|---|---|---|
| LOSS | 53774.69 | $4.04\times10^{10}$ | 10.88 | 5127951 | 11.61 | 194.72 |

Table 7: Descriptive Statistics of the Severity variable

The main reason for choosing the Gamma distribution for this model its good fit to the logarithm of the LOSS variable. Thus, the Gamma distribution with a *log* link was considered in the GLM model. Although it is a common practice to check if the data fits a certain distribution before using it in a model, in the case of Generalized Linear Models (GLMs), the focus is on modeling the conditional expected value of the response variable.

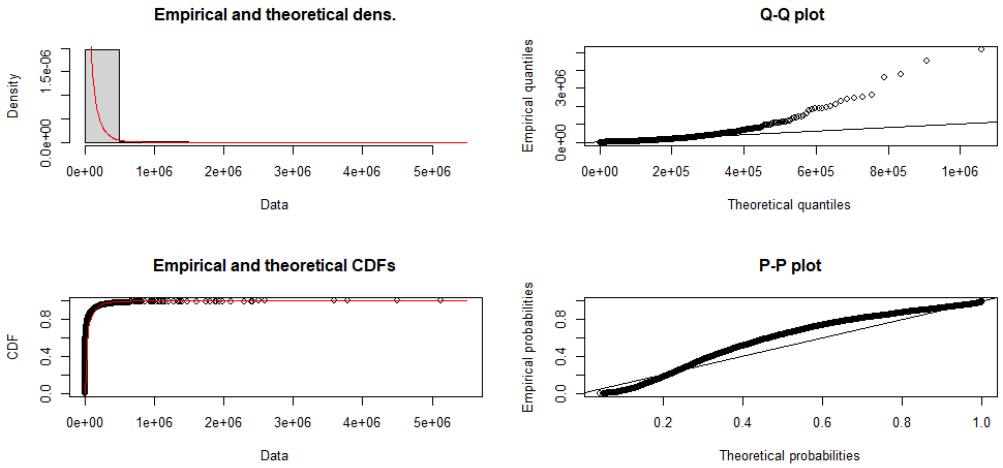


Figure 1: Fit of LOSS to Gamma Distribution

As shown in Figure 1, the original LOSS variable does not exhibit a particularly strong fit to the Gamma distribution. The Q-Q plot highlights significant deviations in the right tail, and the P-P plot suggests a systematic underestimation of probability mass in the intermediate percentiles. These patterns reflect the presence of high-magnitude outliers—a common feature in loss data.

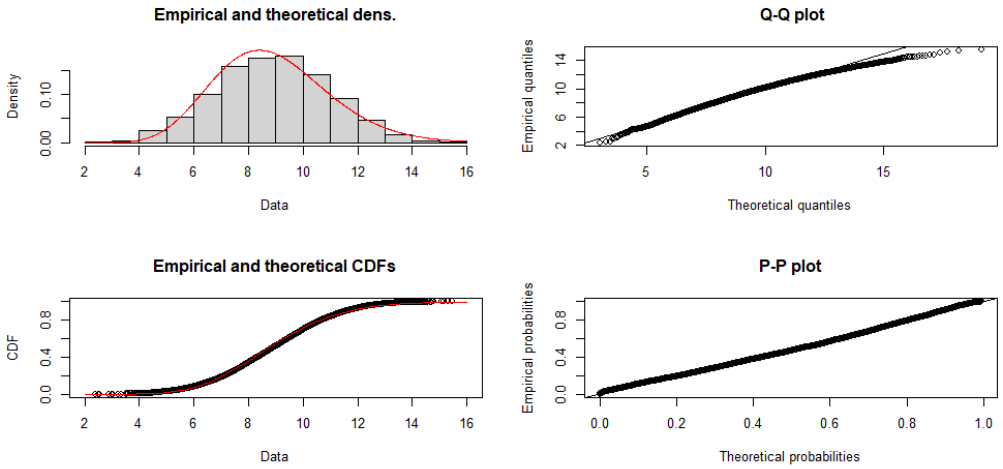


Figure 2: Fit of log(LOSS) to Gamma Distribution

In contrast, the transformed variable log(LOSS), illustrated in Figure 2, displays a markedly improved distributional behavior. The logarithmic transformation yields a more symmetric and regular distribution. The Gamma model aligns well with log(LOSS), as seen in the nearly linear Q-Q plot and the strong agreement between theoretical and empirical densities. Therefore, the good fit of the Gamma distribution to the log(LOSS) variable supports the choice of using a GLM with Gamma distribution and *log* link, even though the original LOSS values do not follow a Gamma distribution exactly.

In the next chapters, we will examine the justification for selecting these variables as response and explanatory variables, as well as explore their statistical behavior in the context of the dataset.

# 4 Methodology

In this chapter, the theoretical framework supporting the development of this dissertation is presented. The discussion covers the fundamental concepts underlying the estimation of insurance premiums, the formulation and components of Generalized Linear Models (GLMs), and the metrics employed to assess the quality of the resulting estimations.

## 4.1 Premium Estimation

The premium charged to the policyholder consists of two main components[1]:

- **Pure Premium** — Represents the fundamental component of the premium, usually reflecting both the expected number of claims occurring in a given time period and the expected average cost of the corresponding indemnity.
- **Safety Loading** — Includes the insurer's operational expenses, the intended profit margin, as well as any taxes and regulatory charges applicable to the insurance contract.

Taking these components into account, the total insurance premium $PT$ is given by[1]:

$$PT = PP + SL$$

where $PP$ denotes the pure premium and $SL$ the safety loading. Sometimes, the safety loading is proportional to the Pure Premium: $SL = \alpha \times PP\ ,\ \alpha > 0$.

This work focuses exclusively on the calculation and estimation of the pure premium, since the value of the Safety Loading is determined and adjusted by the insurer's commercial and financial departments, which falls outside the scope of this Thesis.

Considering each policy within the portfolio, the total claim cost, denoted by $S$, can be given by[1]:

$$S = \sum_{i=1}^{N} X_i \ \ , \tag{2}$$

where $N$ represents the number of claims that occur in a given time period, usually one year, and $X_i$ denotes the cost of the $i$-th claim, with $i = 1, \ldots, N$.

Assuming that the number of claims (frequency) is independent of the individual claim amounts (severity), the expected value of $S$ is[1]:

$$\mathbb{E}[S] = \mathbb{E}\left[\sum_{i=1}^{N} X_i\right] = \mathbb{E}\left[\mathbb{E}\left[\sum_{i=1}^{N} X_i \,\middle|\, N = n\right]\right] = \mathbb{E}[N \cdot \mathbb{E}[X]] = \mathbb{E}[N] \cdot \mathbb{E}[X] \tag{3}$$

Thus, in this context, the *pure premium* corresponds to the expected total claim cost per policy, that is:

$$PP = \mathbb{E}[S] = \mathbb{E}[N] \cdot \mathbb{E}[X]$$

where $\mathbb{E}[N]$ represents the expected number of claims in the time period (frequency), and $\mathbb{E}[X]$ denotes the expected claim amount (severity).

The **Pure Tax** is a key ratio defined as the proportion between the pure premium and the insured capital:

$$PureTax = \frac{PP}{InsuredCapital} = \frac{\mathbb{E}[N] \times \mathbb{E}[X]}{InsuredCapital} \tag{4}$$

### 4.1.1 Frequency

Frequency is represented by the expected number of claims in the time period, denoted by $\mathbb{E}[N]$.

Although the Poisson distribution is commonly used in this type of problem[2], given the dichotomous nature of the variables under consideration (i.e., the occurrence or non-occurrence of a given type of claim), the Bernoulli distribution is adopted here. Specifically, we assume that $N \sim \text{Bernoulli}(p)$, where $p$ denotes the probability of claim occurrence, with $0 \leq p \leq 1$. In this case, the expected number of claims in the time period is the probability of a claim occurring in that period. The probability mass function is given by:

$$f(n) = p^n(1-p)^{1-n} \ , \quad n \in \{0,1\} \ , \quad 0 < p < 1 \ , \tag{5}$$

with

$$\mathbb{E}[N] = p, \quad \text{Var}(N) = p(1-p) \ .$$

Since the Bernoulli distribution is specifically designed to model binary outcomes, it is particularly suitable for representing the occurrence of individual claims in this context. Nevertheless, for other types of frequency modeling involving count data, alternative distributions such as the Poisson or Negative Binomial are frequently employed[2].

### 4.1.2 Severity

Severity is represented by the expected cost of a claim, denoted by $\mathbb{E}[X]$. It is common practice to use the Gamma distribution to model this component[2].

Let $X$ be a continuous random variable. We say that $X$ follows a Gamma distribution, denoted by $X \sim \text{Gamma}(\alpha, \theta)$, with shape parameter $\alpha > 0$ and scale parameter $\theta > 0$. Its probability density function is given by:

$$f(x) = \frac{x^{\alpha-1}}{\Gamma(\alpha)\,\theta^{\alpha}} e^{-\frac{x}{\theta}} \ , \quad x > 0 \ , \quad \alpha, \theta > 0 \ , \tag{6}$$

with

$$\mathbb{E}[X] = \alpha\theta, \quad \text{Var}[X] = \alpha\theta^2 \ .$$

The Gamma distribution is defined only for positive values and exhibits right skewness, making it a suitable choice for modeling claim severity. Nevertheless, alternative distributions are also commonly employed for modeling severity, such as the Lognormal, Pareto, or Weibull distributions. These distributions were not considered, as the response variables in GLM must belong to the exponential family—a condition that these three distributions do not satisfy[2].

## 4.2 GLM

Generalized Linear Models (GLMs) are an extension of classical linear regression models. Through a link function, GLMs relate a linear combination of explanatory variables, $X_1, X_2, \ldots, X_p$, to the expected value of the response variable $Y$[2]. Unlike classical linear models, the response variable in a GLM can follow any distribution from the exponential family (with the Normal distribution as a particular case)[2].

A GLM consists of three fundamental components:

- **Random Component** — The response variable $Y$, which is assumed to follow a distribution belonging to the exponential family[2].
- **Systematic Component** — The explanatory variables, that are combined linearly to form the predictor, as follows[2]:

$$\eta_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_p X_{pi}, \quad i = 1, \ldots, n \tag{7}$$

  where $\eta_i$ is the linear predictor for observation $i$, $\beta_j$ is the model coefficient of explanatory variable $j$, $j = 1, \ldots, p$ , and $X_{ji}$ represents the $j$-th explanatory variable for the $i$-th observation. This can be expressed in matrix notation as:

$$\vec{\eta} = \mathbf{X}\vec{\beta}\,, \tag{8}$$

  where $\vec{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^\top$ is the vector of coefficients, and $\mathbf{X}$ is the design matrix of explanatory variables, defined as:

$$\mathbf{X} = \begin{bmatrix} 1 & X_{11} & X_{21} & \cdots & X_{p1} \\ 1 & X_{12} & X_{22} & \cdots & X_{p2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{1n} & X_{2n} & \cdots & X_{pn} \end{bmatrix}$$

- **Link Function** — A differentiable and monotonic function $g(\cdot)$ that connects the linear predictor $\eta_i$ to the expected value of the response variable $\mu_i = \mathbb{E}[Y_i]$, such that[2]:

$$g(\mu_i) = \eta_i, \quad i = 1, \ldots, n\,. \tag{9}$$

  Its inverse, when it exists, is given by:

$$\mu_i = g^{-1}(\eta_i)\,. \tag{10}$$

The flexibility of GLMs, particularly the ability to model non-normal response distributions through appropriate choices of the link function and variance structure, makes them a widely used framework in actuarial science and insurance pricing in particular, and other applied statistical fields[3].

### 4.2.1 Exponential Family of Distributions

A random variable $Y$ is said to belong to the exponential family if its probability density function (for continuous $Y$) or probability mass function (for discrete $Y$) can be expressed in the following canonical form[3]:

$$f(y \mid \theta, \phi) = \exp\left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right)\,, \tag{11}$$

where $\theta$ and $\phi$ are parameters — with $\theta$ being the canonical (or natural) parameter and $\phi$ the dispersion parameter — and $a(\cdot)$, $b(\cdot)$, and $c(\cdot)$ are known real-valued functions.

As previously discussed, given the nature of the variables under study, we will use the Bernoulli distribution to model frequency and the Gamma distribution to model severity.

The Bernoulli$(p)$ distribution[3], with probability mass funtion (5), belongs to the exponential family with the following components:

- $\theta = \ln\left(\frac{p}{1-p}\right)$
- $\phi = 1$
- $b(\theta) = \ln(1 + e^{\theta})$
- $a(\phi) = 1$
- $c(y, \phi) = 0$

For the Gamma$(\alpha, \beta)$ distribution[3], with probability density function (6), which also belongs to the exponential family, expressed in the canonical form with:

- $\theta = \dfrac{1}{\alpha\beta}$
- $\phi = \dfrac{1}{\alpha}$
- $b(\theta) = \ln\theta$
- $a(\phi) = -\phi$
- $c(y, \phi) = \left(\dfrac{1}{\phi} - 1\right)\ln x - \dfrac{\ln\phi}{\phi} - \ln\Gamma\left(\dfrac{1}{\phi}\right)$

Hence, detailed derivations of these representations are provided in [3].

Given that both the Bernoulli and Gamma distributions can be expressed in the exponential family form, they are suitable choices as response distributions within the GLM framework.

To determine the most appropriate link function for both severity and frequency, the *Akaike Information Criterion* is employed as the primary model selection metric. The $AIC$ is defined as[1]:

$$AIC = 2p - 2 \cdot loglik(\beta) \tag{12}$$

where $loglik(\beta)$ denotes the log-likelihood function of the generalized linear model and $p$ represents the number of parameters estimated.

The $AIC$ is, in summary, a method for estimating the amount of information lost by a given model, where the smaller the information loss, the better the model quality and the lower the AIC value. Consequently, the best model is the one with the lowest $AIC$.

We will see in chapter 5.2 that using this approach, the link functions chosen among the most common ones for Bernoulli and Gamma[3] were the *Log* for severity and the *Logit* and *Probit* for frequency.

## 4.3 Quality of Fitting

Once the estimation of the $\beta$ coefficients for the models is complete, it is essential to assess the quality of the fitted values in relation to the observed values of the response variable. This assessment is typically performed through a residual analysis, where the residuals are defined as:

$$\hat{e_i} = y_i - \hat{y_i} \ , \tag{13}$$

representing the difference between the observed values $y_i$ and the fitted values $\hat{y_i}$.

Another commonly used error measure in GLMs are the *Deviance Residuals*, given by[3]:

$$r_D = \text{sign}(y_i - \hat{y_i}) \sqrt{2 \cdot [l(y_i; y_i) - l(y_i; \hat{y_i})]} \tag{14}$$

where $y_i$ are the observed values, $\hat{y_i}$ are the fitted values and $l(\cdot)$ denotes the log-likelihood function. Deviance residuals are commonly used in GLMs as they provide valuable information for diagnosing the fit of the model and conducting goodness-of-fit tests.

Another commonly used metric to evaluate the goodness of fit of the model is the well-known McFadden's $R^2$[5]:

$$R^2_{\text{McFadden}} = 1 - \frac{\ln L_{\text{model}}}{\ln L_{\text{null}}} \ , \tag{15}$$

where $L_{\text{model}}$ is the likelihood of the fitted model, and $L_{\text{null}}$ is the likelihood of the null model (intercept-only, no predictors). This metric ranges between 0 and 1, with values closer to 1 indicating better fit. Values between 0.2 and 0.4 are generally considered excellent for GLM models[5].

# 5 Results

This chapter is divided into several stages, from data preparation to the presentation of the final results. All computations throughout the project were performed using R software with packages *readxl*, *MASS*, *ggplot2*, *dplyr*, *fitdistrplus* and *pscl*.

## 5.1 Data Preparation

As explained in chapter 3, we use the Bernoulli distribution to model frequency and the Gamma distribution to model severity in the GLM. Before fitting the models, it was necessary to ensure that the variables `Municipality` and `Culture` were properly set as categorical factors in R. Likewise, the dichotomous variables `FLE`, `FST`, `HAI`, etc., representing the expected frequency of each event, were also converted into factors.

Regarding the `SUMINS` variable (representing the insured capital), it was necessary to apply a logarithmic transformation to mitigate the large disparities in magnitude between observations. Without this transformation, the numerical method for the maximum likelihood estimate approximation struggled to converge, failing to find a stable set of parameters due to scale discrepancies.

## 5.2 Model Selection

We now assess which link function is more appropriate for each model, based on the AIC values.

### Severity Model

The choice of the link function for the severity model is mostly supported by the reasoning provided in Section 3.3. Nonetheless, another link function was tested — the *inverse* link — but using it resulted in the failure of the GLM estimation procedure. On the other hand, the *log* link worked without any issues, and the numerical method for the GLM parameters estimates converged successfully. Thus, we selected the *log* link function for the severity model.

### Frequency Model

For the frequency models, three link functions were considered viable candidates: *logit*, *probit*, and *cloglog*. Based on the comparative AIC values presented in Appendix 8 and discussed in Chapter 4.2, the choices in Table 8 were made.

| Variable | Name | Link Function |
|---|---|---|
| FLE | Fire, Lightning, Explosion | Logit |
| FST | Frost | Logit |
| HAI | Hail | Logit |
| OTH | Other | Logit |
| PR | Persistent Rain | Logit |
| SNW | Snow | Probit |
| SUN | Sunscald | Logit |
| TOR | Tornado | Logit |
| WTR | Waterspout | Probit |

Table 8: Chosen Link Functions for Frequency Models

## 5.3 Model Estimation

Following the selection of link functions, we proceeded to estimate the $\beta$ coefficients for each GLM model. Using R, we obtained the estimated values for all coefficients associated with the systematic component of each response variable. These coefficients are critical for making predictions, which will be tested in the next chapter. Due to the fact that the estimated $\beta$ coefficients exceeded 350 in total (including 1 intercept, 89 for Culture, 267 for Municipality, and 1 for log_Sumins), these estimates were omitted from the document for the sake of conciseness.

## 5.4 Model Testing

The severity and frequency models were evaluated separately, rather than estimating the pure premium ($PP$) directly. This is due to the lack of actual pure premium values in the dataset, which prevents a direct comparison between predicted and observed premiums. We therefore begin by assessing the severity model, followed by the frequency models.

### Severity Model

The severity model was evaluated by comparing the observed values of LOSS with the fitted values produced by the GLM.

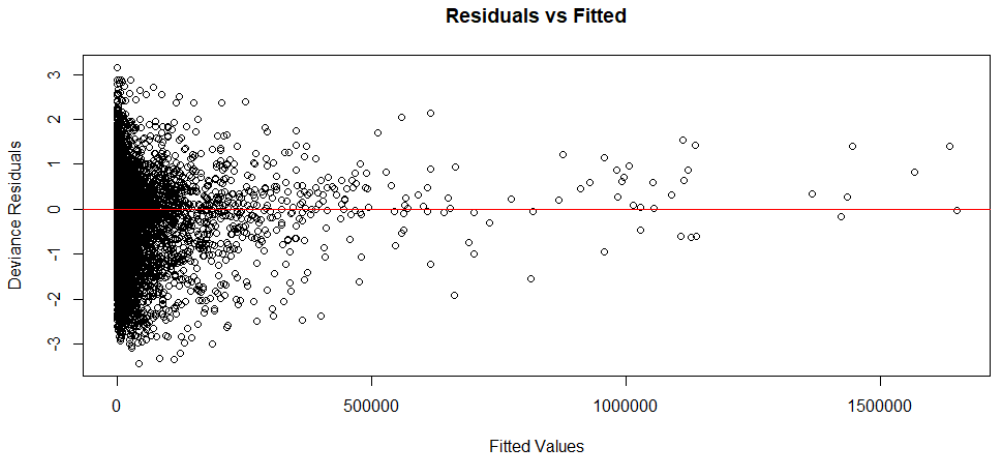


Figure 3: Residuals vs. Fitted Values for the Severity GLM

Figure 3 shows that the deviance residuals are approximately symmetric around zero, indicating that the model does not systematically overestimate or underestimate the response. However, the funnel-shaped pattern suggests a reduction in the spread of residuals as the fitted claim amount values increase. This is indicative of heteroscedasticity and may suggest the presence of a missing explanatory variable or the need for a transformation.

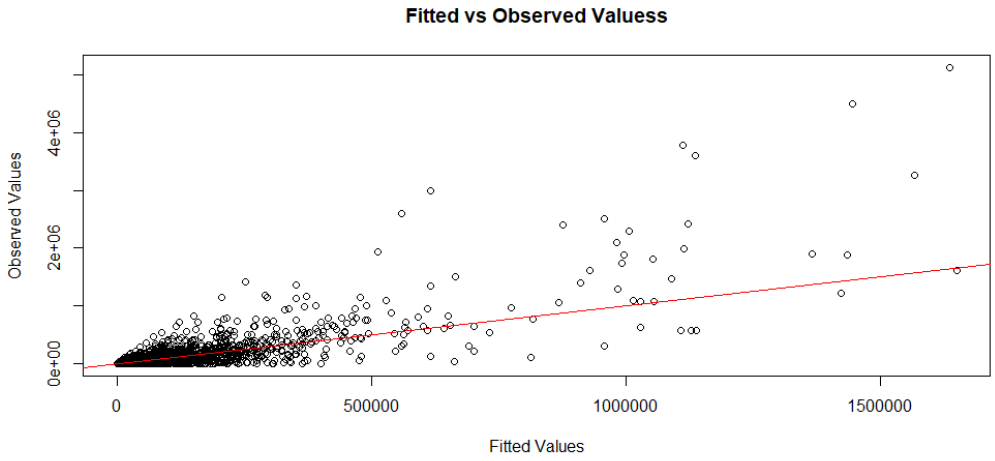


Figure 4: Fitted vs. Observed Values for the Severity GLM

Figure 4 shows that many observed values are substantially higher than the fitted values, revealing that the model tends to underestimate high-severity losses. This is also supported by the fact that the red trend line lies below the identity line, reinforcing the possibility that a relevant explanatory variable may be missing.

In addition, the McFadden's $R^2$ for the severity model is 0.103, indicating a modest improvement over the null model.

It is important to note that, unlike the frequency model, no separate severity model was developed for each type of claim. This decision was due to the unavailability of claim data disaggregated by loss type—only the total aggregated loss (LOSS) is provided, without indication of which portion corresponds to each specific peril.

### Frequency Models

For the frequency component, GLMs were used to estimate the probability of claim occurrence for different risk types. Since the response variable is binary (0 = no claim, 1 = claim), the fitted values from the models represent the estimated probabilities of occurrence for each event.

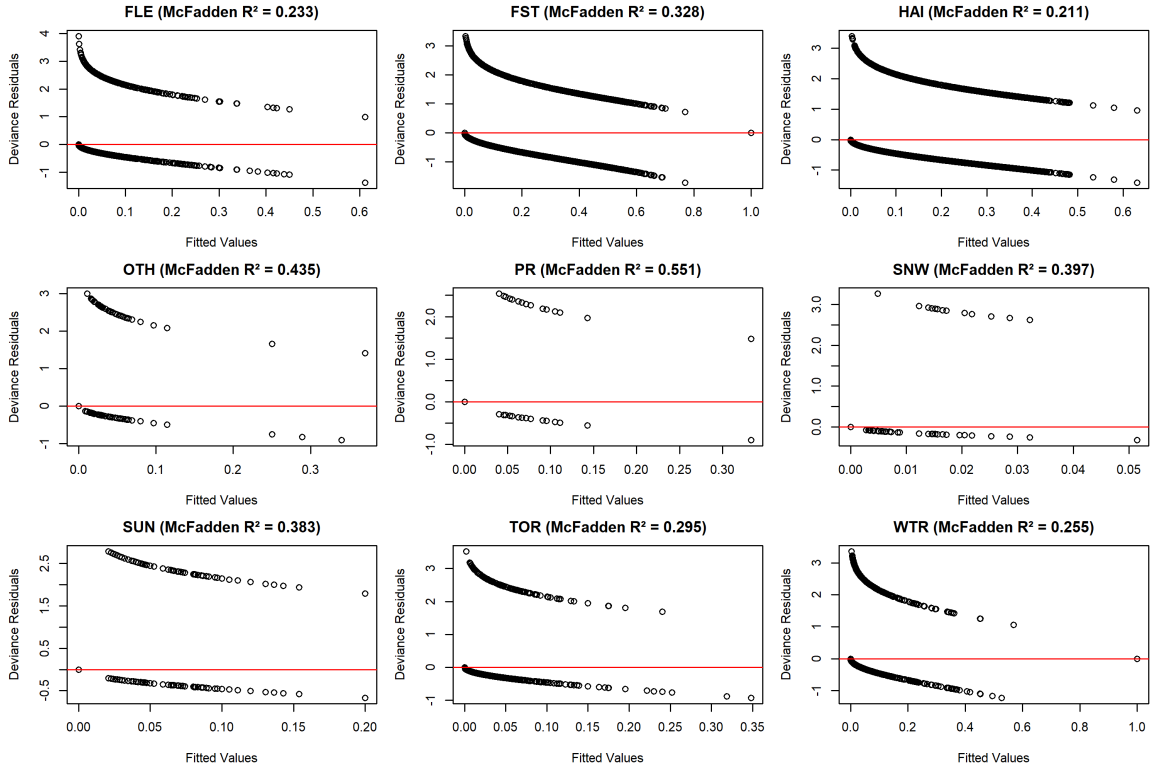


Figure 5: Residuals vs. Fitted Values for the Frequency GLMs

Figure 5 displays curved patterns in the residuals, which are typical in binary models where deviance residuals are inherently asymmetric. The figure also includes McFadden's $R^2$ values for each frequency model, ranging from 0.211 (HAI) to 0.551 (PR). These values suggest a substantial improvement over the null models and indicate that the frequency models perform well in predicting the occurrence of claims.

## 5.5 Company vs. GLM Pure Premium Comparison

We now turn to one of the central parts of this work: the comparison between the pure premium calculated by the company and the one estimated using the GLM approach.

First, a table was created containing all combinations of `Culture` and `Municipality` from the original dataset (as presented in Section 3.1), along with the total insured capital (SUMINS) and total losses (LOSS) for each combination across all observed years. The table was filtered to include only records where LOSS was greater than zero, since zero loss implies that no claim occurred.

For testing purposes, the pure premium was forecasted for a new insurance contract by setting a fixed insured capital value (CAPSEGNEW) of €100 for all `Culture` and `Municipality` combinations, as previously explained. This was done because, in both methods, the insured capital directly affects the premium value. Using different capital values would make it harder to determine

whether differences in the premiums are due to the models themselves or simply due to the size of the insured capital.

| | CULTURE | MUNICIPALITY | CAPSEGNEW | SUMINS | LOSS | log_CAPSEG | PREMIUMCOMP |
|---|---|---|---|---|---|---|---|
| 1 | ABACATE | ALBUFEIRA | 100 | 276915.28 | 21860.28 | 4.60517 | 7.89421236 |
| 2 | ABACATE | FARO | 100 | 1068285.63 | 2505.52 | 4.60517 | 0.23453653 |
| 3 | ABÓBORA | SANTAREM | 100 | 885682.93 | 27622.10 | 4.60517 | 3.11873461 |
| 4 | ALHO | ELVAS | 100 | 342000.00 | 15120.00 | 4.60517 | 4.42105263 |
| 5 | ALPERCE | ESTREMOZ | 100 | 55663.20 | 2637.12 | 4.60517 | 4.73763636 |
| 6 | ALPERCE | FUNDAO | 100 | 6992.00 | 3621.86 | 4.60517 | 51.80005721 |
| 7 | ALPERCE | GUARDA | 100 | 1784.00 | 1052.56 | 4.60517 | 59.00000000 |
| 8 | AMEIXA | ALCOBACA | 100 | 1308579.22 | 93925.71 | 4.60517 | 7.17768618 |
| 9 | AMEIXA | ALENQUER | 100 | 1647563.69 | 72626.08 | 4.60517 | 4.40808938 |
| 10 | AMEIXA | ALFANDEGA DA FE | 100 | 12480.00 | 2620.80 | 4.60517 | 21.00000000 |
| 11 | AMEIXA | ARMAMAR | 100 | 534189.65 | 121969.94 | 4.60517 | 22.83270370 |
| 12 | AMEIXA | BELMONTE | 100 | 222138.17 | 64613.98 | 4.60517 | 29.08729283 |
| 13 | AMEIXA | BOMBARRAL | 100 | 3563589.68 | 158109.12 | 4.60517 | 4.43679363 |
| 14 | AMEIXA | BORBA | 100 | 2847154.07 | 58388.81 | 4.60517 | 2.05077802 |
| 15 | AMEIXA | BRAGA | 100 | 7591.45 | 1853.13 | 4.60517 | 24.41075157 |
| 16 | AMEIXA | CADAVAL | 100 | 1518927.93 | 41505.91 | 4.60517 | 2.73257929 |
| 17 | AMEIXA | CALDAS DA RAINHA | 100 | 162598.67 | 24838.36 | 4.60517 | 15.27586864 |
| 18 | AMEIXA | CAMPO MAIOR | 100 | 20693424.85 | 1937935.65 | 4.60517 | 9.36498267 |
| 19 | AMEIXA | CARRAZEDA DE ANSIAES | 100 | 42482.14 | 9325.77 | 4.60517 | 21.95221333 |

Table 9: Aggregated Table by Municipality and Culture

The company premium, referred to here as PREMIUMCOMP, was calculated using the method currently employed by the company, as described in Equation (1). In practice, the pure premium rate iscurrently computed as:

$$\text{PureTax} = \frac{\text{LOSS}}{\text{SUMINS}},$$

and the pure premium is then given by:

$$\text{PREMIUMCOMP} = \text{PureTax} \times \text{CAPSEGNEW}.$$

The GLM-based pure premium was obtained by combining the frequency and severity models. For the frequency component, we summed the expected number of claims $\mathbb{E}[N]$ across all nine models—one for each type of claim. This is necessary because the IFAP dataset provides aggregated losses that include all types of claims. The final premium, $\mathbb{E}[S]$, was then calculated by multiplying the total expected frequency by the expected severity $\mathbb{E}[X]$ predicted by the severity model:

$$PP = \mathbb{E}[S] = \mathbb{E}[N] \times \mathbb{E}[X] = \left( \sum_{i=1}^{9} \mathbb{E}[N_i] \right) \mathbb{E}[X].$$

where $\mathbb{E}[N_i]$ is the expected number of claims for each type of claim $i = 1, \ldots, 9$.

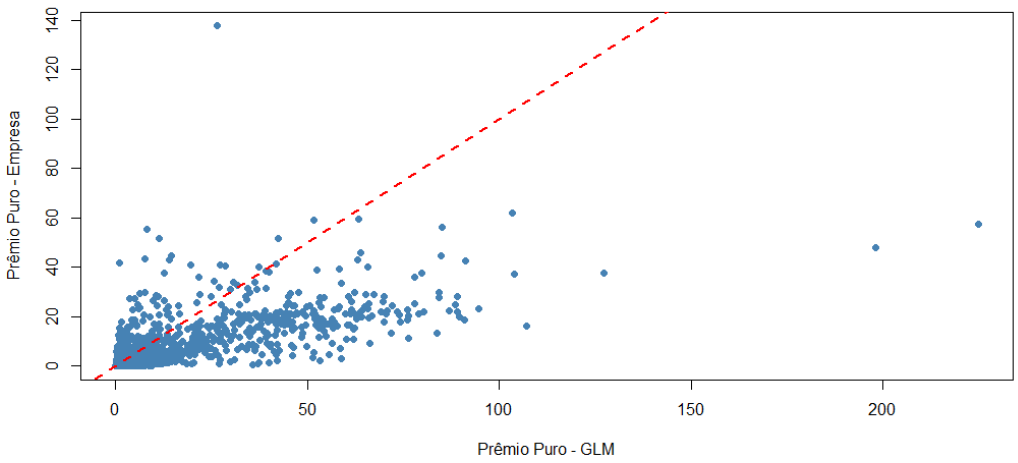


Figure 6: Pure Premium: GLM vs. Company Method (Linear Scale)

As shown in Figure 6, most of the $(PP_{GLM}, PP_{Company})$ points, where $PP_{GLM}$ is the Pure Premium estimate by GLM and $PP_{Company}$ by company, correspond to low premium values under both methods. This high concentration of values around the origin makes visual comparison difficult. To improve interpretability, we applied a logarithmic scale to both axes.

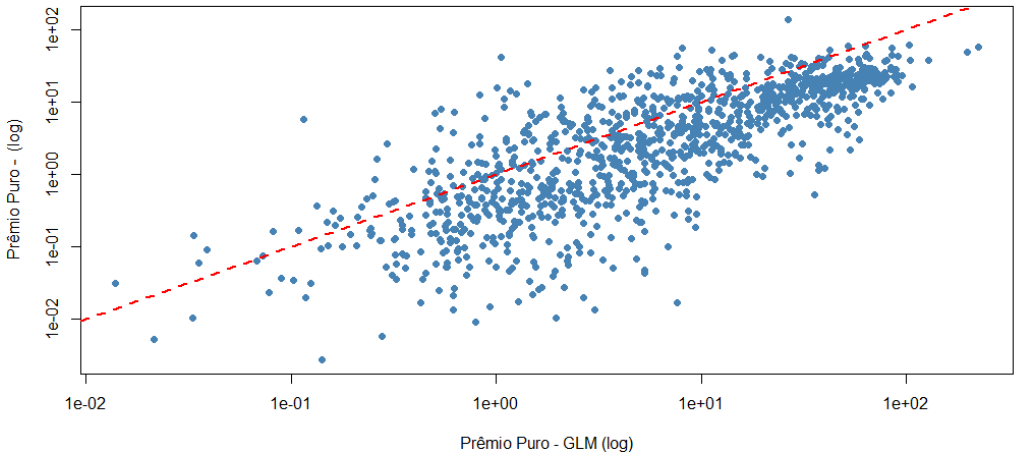


Figure 7: Pure Premium: GLM vs. Company Method (Log Scale)

From Figure 7, several remarks can be made. The points generally follow a linear trend, indicating a strong proportional relationship between the two methods. However, for lower premium values (toward the left side of the graph), there is greater dispersion, suggesting less agreement between the models in this range.

The red line in the figure represents perfect equality between $PP_{GLM}$ and $PP_{Company}$. Points above the line indicate that the company charges more than the GLM suggests, while points below indicate the opposite. Notably, most points fall below the line, meaning the GLM tends to propose a higher premium than the company currently charges. This may indicate that the company's current pricing model could be underestimating the true risk in certain municipality-crop combinations, suggesting a potential need for model revision.

# 6 Conclusion

The research developed in this work aimed to improve the pure premium pricing process in agricultural insurance, focusing on the comparison between the traditional model used by Atlas MGA and the application of Generalized Linear Models (GLMs). This topic is highly relevant, as ensuring an accurate estimation of the pure premium is crucial both for the insurer and the insured.

The application of the GLM proved to be a robust and efficient alternative for modeling the pure premium through the expected claim frequency and severity. Firstly, the model was successfully developed and implemented using relevant variables for this study, such as crop type, municipality, and the type of weather-related event. Secondly, the comparison between the pure premium estimated using the GLM and the one calculated through the approach employed by Atlas MGA showed that the GLM was capable of providing accurate and robust predictions. Moreover, the results suggest that the traditional method may underestimate the pure premium in certain cases, which could imply financial risks for both insurers and policyholders.

This project not only deepened the understanding of the application of GLMs in agricultural insurance but also proposed potential improvements to the approach currently used by Atlas MGA, providing a foundation for future innovation in this field.

As future contributions, the developed models could be continuously applied to updated claims databases, as well as to new climatic trends, with the aim of designing new insurance coverages. It is also important to explore and test new explanatory variables that may improve model accuracy and enhance pure premium prediction. Additionally, further research into alternative pricing methods—such as neural networks or machine learning techniques—could contribute to the ongoing refinement of the premium estimation process.

# 7 Bibliography


[1] Klugman, S. A., Panjer, H. H., & Willmot, G. E. (2019). Loss Models: From Data to Decisions (5th ed.). Wiley.

[2] Ohlsson, E., & Johansson, B. (2015). Non-Life Insurance Pricing with Generalized Linear Models (3rd ed.). Springer.

[3] McCullagh, P., & Nelder, J. A. (1989). Generalized Linear Models (2nd ed.). Chapman and Hall.

[4] Hardin, J. W., & Hilbe, J. M. (2007). Generalized Linear Models and Extensions (2nd ed.). Stata Press.

[5] McFadden, D. (1978). Quantitative methods for analyzing travel behavior of individuals: Some recent developments. In D. A. Hensher & P. R. Stopher (Eds.), Behavioral Travel Modeling. London: Croom Helm.

[6] Goldburd, M., Khare, A., Tevet, D., & Guller, D. (2025). Generalized Linear Models for Insurance Rating (2nd ed., revised). Casualty Actuarial Society.

[7] Sebyhed, H. (n.d.). Machine Insurance Premium Calculations Based on Claim Models [Master's thesis]. [Institution not specified].
link: https://www.diva-portal.org/smash/get/diva2:1792576/FULLTEXT01.pdf

[8] Cadima, J. (2021). Modelos Matemáticos e Aplicações: Modelos Lineares Generalizados [Slides]. Instituto Superior de Agronomia, Universidade de Lisboa.
link: https://fenix.isa.ulisboa.pt/downloadFile/281547991169344/acetatosMLG.pdf

# 8 Appendix

## Link Functions

| **Link Funtion** | Logit | Probit | Cloglog |
|---|---|---|---|
| **AIC** | 5890.3 | 5890.6 | 5895.3 |

Table 10: Link Functions AIC for FLE

| **Link Funtion** | Logit | Probit | Cloglog |
|---|---|---|---|
| **AIC** | 13685 | 13690 | 13685 |

Table 11: Link Functions AIC for FST

| **Link Funtion** | Logit | Probit | Cloglog |
|---|---|---|---|
| **AIC** | 15931 | 15936 | 15934 |

Table 12: Link Functions AIC for HAI

| **Link Funtion** | Logit | Probit | Cloglog |
|---|---|---|---|
| **AIC** | 1230.3 | 1230.8 | 1230.8 |

Table 13: Link Functions AIC for OTH

| **Link Funtion** | Logit | Probit | Cloglog |
|---|---|---|---|
| **AIC** | 978.13 | 978.13 | 978.13 |

Table 14: Link Functions AIC for PR

| **Link Funtion** | Logit | Probit | Cloglog |
|---|---|---|---|
| **AIC** | 897.35 | 896.95 | 897.38 |

Table 15: Link Functions AIC for SNW

| **Link Funtion** | Logit | Probit | Cloglog |
|---|---|---|---|
| **AIC** | 2090.1 | 2090.1 | 2090.1 |

Table 16: Link Functions AIC for SUN

| **Link Funtion** | Logit | Probit | Cloglog |
|---|---|---|---|
| **AIC** | 2652 | 2652.4 | 2652.2 |

Table 17: Link Functions AIC for TOR

| **Link Funtion** | Logit | Probit | Cloglog |
|---|---|---|---|
| **AIC** | 3892.3 | 3883.7 | 3895.5 |

Table 18: Link Functions AIC for WTR