



Università degli Studi di Trieste

Dipartimento di Scienze Economiche, Aziendali, Matematiche e Statistiche
Corso di Laurea Magistrale in Scienze Statistiche e Attuariali

TESI DI LAUREA

Application of GLM Advancements to Non Life Insurance Pricing

Candidato:

Leonardo Stincone

Relatore:

Prof. Francesco Pauli

*The data scientist is a person
who is better at statistics than any software engineer
and better at software engineering than any statistician.*

Josh Wills

Acknowledgements

Ringraziamenti vari

Abstract

This is my abstract ...

Table of Contents

List of Figures	viii
List of Tables	ix
Introduction	1
Thesis aim	1
Actuary and datascientist figure	1
Thesis structure	1
1 Non-Life Insurance Pricing	2
1.1 What a Non-Life Insurance is	2
1.2 Non-Life insurance pricing	5
1.2.1 Compound distribution hypotheses	5
1.2.2 Distribution of the total cost of claims	6
1.2.3 Risk premium and Technical Price	8
1.3 Modeling and Personalization	9
1.3.1 Pricing variables	9
1.3.2 Pricing variables encoding	11
1.3.3 Pricing Rule and Modeling	13
1.3.4 Response variables and distributions	13
1.3.4.1 Distribution for the claims count N	14
1.3.4.2 Exposure	15
1.3.4.3 Distribution for the claim severity Z	18
1.3.4.4 Large claims	21
1.3.4.5 Binomial distribution	24
1.3.5 Model fitting and data available	27
1.3.5.1 Settlement process and IBNR	29

TABLE OF CONTENTS

1.3.5.2	Model fitting with available data	30
1.4	Beyond technical pricing	33
1.4.1	Tariff and Offer Price	34
1.4.2	Commercial price	35
1.5	The actuary role	37
2	Statistical models for Non Life Insurance Pricing	39
2.1	Statistical Models	39
2.1.1	GLM	40
2.1.1.1	Linear Exponential Family	40
2.1.1.2	Model assumptions	42
2.1.1.3	Model fitting	44
2.1.1.4	Variable effects	47
2.1.1.5	Variable selection	47
2.1.2	Elastic Net	48
2.1.3	GAM	48
2.1.4	GBM	49
2.2	Model comparison	50
2.3	The actuary importance	50
2.4	Implementation	51
3	Practical application	53
3.1	Data description	53
3.2	Model used	54
3.3	Model assessment	55
3.4	Results	55
	Bibliography	57

List of Figures

1.1	Insurance Contract cash flows.	4
1.2	Poisson distribution for some values of λ	14
1.3	Gamma distribution for some values of α and ρ	19
1.4	Large claims.	22
1.5	Binomial distribution for some values of n and p	25
1.6	Claim timeline.	30
2.1	Design Matrix \mathbf{X}	43

List of Tables

1.1	Dummy variables encoding.	12
2.1	Some Linear Exponential Families.	41
2.2	Canonical link functions.	44
2.3	Deviance for Linear Exponential Families	47

Introduction

La mia introduzione ...

Thesis aim

Lorem ipsum ...

Actuary and datascientist figure

Lorem ipsum ...

Thesis structure

Lorem ipsum ...

→

1

Non-Life Insurance Pricing

In this chapter I am going to provide an overview on how non-life insurance works from an actuarial point of view with a specific focus on the retail pricing process.

1.1 What a Non-Life Insurance is

The Italian Civil Code provides the following definition of insurance contract:

Definition 1.1 (Insurance Contract, Art. 1882, Italian Civil Code). The insurance is the contract by which an insurer, in exchange of the payment of a certain premium, obliged himself, within the agreed limits:

1. to pay an indemnity to the insured equivalent to the damage caused by an accident;
2. or to pay an income or a capital if a life-related event occurs.

This definition identifies two parties: the *Insurer* and the *Policyholder*. The policyholder pays to the Insurer a certain *Premium* at the beginning of the insurance coverage and the insurer will pay a benefit if a certain event (*Claim*) occurs. This event could happen zero, one or more than one times, so it is possible to have more than one claim.

Usually, in non-life insurance, the benefit is the payment of a sum. This sum could be predetermined (e.g. in motor theft insurance, where the benefit is usually the value of the insured vehicle) or defined by the entity of the claim (e.g. in motor third party liability insurance, it depends on the damage the policyholder has provided to a third party). Regarding the “agreed limits”, another peculiarity of non-life insurances is that the coverage period is defined as a fixed amount of time, usually corresponding to 1 year.

Starting from this legal definition, we can formalize a non-life insurance contract as follows.

Let’s:

- $]t_1, t_2]$, with $t_1 < t_2$, be the coverage period;
- $P > 0$ be the premium payed by the policyholder to the insurer;
- $N \in \mathbb{N}$ be the number of claims occurred during the coverage period (*claims count*);
- $\tau_1, \tau_2, \dots, \tau_N$, with $t_1 < \tau_1 < \tau_2 < \dots < \tau_N < t_2$, be the timing of each claim;
- $Z_1, Z_2, \dots, Z_N > 0$ be the amount of each claim (*claims severities* or *claims sizes*).

The total cost of claims for the insurance is:

$$S = \begin{cases} 0 & \text{if } N = 0 \\ \sum_{i=1}^N Z_i & \text{if } N > 0 \end{cases}$$

For semplicity, in the following we are going to just use the notation $S = \sum_{i=1}^N Z_i$ with the meaning of 0 if $N = 0$.

Figure 1.1 shows the cash flows corresponding to the insurance contract. From this representation we can interpret the entering into an insurance contract by the policyholder as a way to exchange the negative cash flows $-Z_1, -Z_2, \dots, -Z_N$ with one single negative cash flow $-P$. On the other hand, the insurer undertakes the negative cash flows $-Z_1, -Z_2, \dots, -Z_N$ in exchange for a positive cash flow $+P$.

The major difference between these cash flows is that P is a certain amount, while Z_1, Z_2, \dots, Z_N , at the time t_1 , are uncertain in the amount, in the count (N)

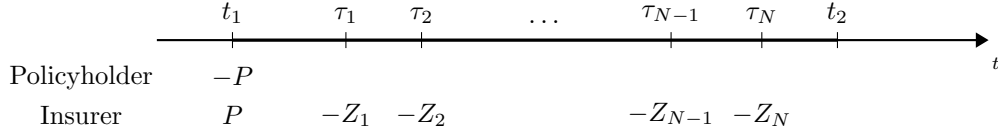


Figure 1.1: Insurance Contract cash flows.

and in the timing $(\tau_1, \tau_2, \dots, \tau_N)$. So, the policyholder, paying a premium P , is giving his risk to the insurer.

This representation points out the inversion of the production cycle typical of the insurance activity. From the insurer point of view, the revenue emerges at the beginning of the economic activity, in t_1 , while the costs will emerge later. In most of other economic activities, the costs emerge before the selling of the product, so the agent can choose the selling price taking into account how much that product costed him. In insurance activity, the insurer, when is selling his product (the insurance coverage), doesn't know the amount of costs he is going to pay for that product. Thus, it is crucial to properly estimate the future costs and determine an adequate premium.

From a statistical point of view, we can translate this uncertainty saying that N and Z_1, Z_2, \dots, Z_N are random variables. Therefore, we can say that $\{N, Z_1, Z_2, \dots\}$ is a stochastic process. Usually, in non-file insurance pricing, the variables $\tau_1, \tau_2, \dots, \tau_N$ are not taken into account because the coverage span is short and from a financial point of view the timing of the claims occurrences is negligible.

Previously we said that Z_1, Z_2, \dots, Z_N are all > 0 . This assumption corresponds to the fact that we are excluding the null claims, i.e. the claims that have been opened, but result in no payment due by the insurer. For the values of Z_i with $N < i$ we can use the rule that $\{N < i\} \Rightarrow \{Z_i = 0\}$, so $Z_{N+1} = 0, Z_{N+2} = 0, \dots$. Therefore, we can say that:

$$\{N < i\} \iff \{Z_i = 0\}$$

1.2 Non-Life insurance pricing

In insurances, the premium that the insurer offers to the policyholder in exchange for the insurance coverage is not the same for every policyholder. The insurer evaluates the risk related to that policy and determine a “proper” premium taking into account risk related factors and commercial related factors. The process of *pricing* corresponds in defining the set of rules for determining this “proper” premium P_i for a specific policyholder i , given the known information on him. In the next sections I am going to better explain what “proper” means.

1.2.1 Compound distribution hypotheses

The first step for evaluating the stochastic process $\{N, Z_1, Z_2, \dots\}$ is to introduce some probabilistic hypotheses. The usual hypotheses assumed are the following:

Definition 1.2 (Compound distribution). Let’s assume that:

1. for each $n > 0$, the variables $Z_1|N = n, Z_2|N = n, \dots, Z_n|N = n$ are stochastically independent and identically distributed;
2. the probability distribution of $Z_i|N = n, i \leq n$ does not depend on n .

Under these hypotheses we say that:

$$S = \sum_{i=1}^N Z_i$$

has a compound distribution.

The variable $Z_i|N = n$ used in this definition can be interpreted as the *claim severity for the i^{th} claim under the hypothesis that n claims occurred*. The two hypotheses provided in definition 1.2 imply that the distribution of $Z_i|N = n, i < n$ does not depend from i nor from n . For this reason, in the following, we are going to use the notation Z to represent a random variable with the $Z_i|N = n, i < n$ distribution and $F_Z(\cdot)$ for its cumulative distribution function (i.e. $F_Z(z) = P(Z \leq z)$).

Let’s consider the variable $Z_i|N > i$. We can interpret it as the *claim severity for the i^{th} claim under the hypothesis that the i^{th} claim occurred*. From the hypotheses

provided in definition 1.2 we can obtain that also $Z_i|N > i$ has the same distribution of $Z_i|N = n$, $i < n$. This can be easily obtained as follows:

$$P(Z_i|N \geq i) = P\left(Z_i \middle| \bigvee_{n=i}^{+\infty} (N = n)\right) \quad (1.1)$$

$$= \sum_{n=i}^{+\infty} \underbrace{P(Z_i \leq z|N = n)}_{=F_Z(z)} P(N = n|N \geq i) \quad (1.2)$$

$$= \sum_{n=i}^{+\infty} F_Z(z) P(N = n|N \geq i) \quad (1.3)$$

$$= F_Z(z) \underbrace{\sum_{n=i}^{+\infty} P(N = n|N \geq i)}_{=1} \quad (1.4)$$

$$= F_Z(z)$$

Where:

- the step (1.1) and the step (1.2) are given by the fact that the event $\{N \geq i\}$ can be decomposed as $\{N \geq i\} = \{\bigvee_{n=i}^{+\infty} (N = n)\}$ and that the events $\{N = n\}, n \in \{i, i+1, i+2, \dots\}$ are two-by-two disjoint, so they constitute a partition of $\{N \geq i\}$, that allows us to use the disintegrability property of the probability;
- the step (1.3) is due to the fact that the distribution of $Z_i \leq z|N = n$ depends neither on i nor on n ;
- the equivalence $\sum_{n=i}^{+\infty} P(N = n|N \geq i) = 1$ at step (1.4) is due to the fact that the events $\{N = n\}, n \in \{i, i+1, i+2, \dots\}$ are a partition of $\{N \geq i\}$.

Therefore, Z can be considered as the *claim severity for a claim under the hypothesis that that claim occurred*.

1.2.2 Distribution of the total cost of claims

Under the hypotheses defined in definition 1.2, it is possible to obtain the full distribution of S given the distribution of N and Z . In this chapter we are going to

provide only the formula of the expected value $E(S)$, but, with the same approach one can obtain all the moments.

The expected value of the total cost of claims $E(S)$ can be obtained from the expected value of the claims count $E(N)$ and the expected value of the claim severity $E(Z)$ as follows:

$$E(S) = \sum_{n=0}^{+\infty} P(N = n) E(S|N = n) \quad (1.5)$$

$$= \sum_{n=0}^{+\infty} P(N = n) E\left(\sum_{i=1}^n Z_i \middle| N = n\right) \quad (1.6)$$

$$= \sum_{n=0}^{+\infty} P(N = n) \sum_{i=1}^n \underbrace{E(Z_i|N = n)}_{=E(Z)} \quad (1.7)$$

$$= \sum_{n=0}^{+\infty} P(N = n) n E(Z) \quad (1.8)$$

$$= E(Z) \underbrace{\sum_{n=0}^{+\infty} n P(N = n)}_{=E(N)} \quad (1.9)$$

$$= E(N) E(Z) \quad (1.10)$$

Where:

- the step (1.5) is given by the fact that the events $\{N = 0\}, \{N = 1\}, \{N = 2\}, \dots$ constitute a partition of the certain event Ω , that allows us to use the disintegrability property of the expected value;
- the step (1.6) is due to the definition of S ;
- the step (1.7) is due to the linearity of the expected value;
- the steps (1.8) and (1.9) are due to the fact that, as assumed by the compound distribution hypotheses, $E(Z_i|N = n)$ does not depends on i and n ;
- the step (1.10) is due to the definition of the expected value $E(N) = \sum_{n=0}^{+\infty} n P(N = n)$.

This result tells us that, under the hypotheses of the compound distribution, it is possible to easily obtain $E(S)$ from $E(N)$ and $E(Z)$. That means that we

can model separately $E(N)$ and $E(Z)$ and, from them, obtain $E(S)$. That result is particularly useful in personalization (paragraph 1.3), because, for each individual i , given the information we have on him, we can estimate his expected claim size $E(N_i)$ and his expected claim severity $E(Z_i)$ and obtain his expected total cost of claims as $E(S_i) = E(N_i)E(Z_i)$.

1.2.3 Risk premium and Technical Price

The expected cost of claims $E(S)$ is important because it gives us a first interpretation of what “proper” premium means.

Definition 1.3 (Risk Premium). Said S the total cost of claims of a policyholder, his *Risk Premium* is given by:

$$P^{(risk)} = E(S)$$

The *Risk Premium* is the premium that on average covers the total cost of claims. As mentioned above, as the coverage spans are usually short, we are not taking into account the timing of the claims so we don’t discount the fact that the claims occur later than the premium payment.

It is clear that this premium, that only covers the cost of claims, is not “proper” in the practice.

First of all, the insurer has to cover also the expenses related to the policy (commission on sales and expenses related to the claim settlement) and the general expenses of the company. Adding the expenses, we obtain the *Technical Price*.

Definition 1.4 (Technical Price). Said S the total cost of claims of a policyholder and E the expenses related to his policy, his *Technical Price* is given by:

$$P^{(tech)} = E(S) + E = P^{(risk)} + E$$

Secondly, even if the policyholder would pay a premium that on average covers claims and expenses, undertaking that risk with nothing in return would not make

sense for the insurer. So, to the Technical Price, some further loadings must be added, as for example risk margin and profit margin.

The result of the Technical Price with these loadings can be further modified based on business logic, as I am going to discuss later.

1.3 Modeling and Personalization

In this section we are going to better explain how pricing based on policyholder information works.

1.3.1 Pricing variables

Usually for every policyholder we have a certain amount of information on him that is considered relevant for his risk evaluation. This information must be reliable and observable at the moment of the underwriting of the policy.

In motor insurances, this information could be:

- Information on the insured vehicle: make, model, engine power, vehicle mass, age of the vehicle;
- General information of the policyholder: age, sex, address (region, city, postcode), ownership of a private box where he parks the car;
- Insurance specific information of the policyholder: number of claims caused in the previous years, how long he has been covered, bonus-malus class;
- Policy options: amount of the maximum coverage, presence and amount of a deductible, presence of other insurance guarantees, how many drivers will drive the vehicle;
- Customer information on the policyholder: how many years he has been a customer of the insurer, how many other policies he owns.
- Telematic data: how many kilometers per year the policyholder travelled in the previous years, how many sharp accelerations and decelerations per kilometer the policyholder performed in the previous years.

These pieces of information are usually called *pricing variables*.

We must observe that some of these variables are available for every potential customer (such as his age and address), while others are only available for policyholder that are already customers (such as telematic data that is available only if the policyholder agreed on installing on their car the device that collects this data).

Moreover, even considering the variables that are available for every customer, it is important to be aware on how reliable they are. Some of them comes from official documents (as customer age and address or bonus-malus class), but others could be declared by the customer and his statements are not easily verifiable by the insurer (as the ownership of a private box or how many drivers will drive the vehicle).

This topic of variables reliability fits in the wider framework of fraud detection. Insurance companies put a lot of effort in preventing frauds. This is done with active actions, such as documents checks and inspections, and with predictive fraud detections models. The two most common categories of frauds are underwriting frauds (such as false declaration on insurance related data) and settlement frauds (such as faking an accident). The customer information on the policyholder is usually important to predict both underwriting frauds and settlement frauds. Usually customers that have a longer relationship with the company and own many policies are less likely to commit frauds.

Regarding the topic of variables reliability, the Italian Insurance Associations ([ANIA](#)) in the last years made some big steps forward by collecting in its databases a lot of information about policyholders and vehicles and making it available to insurance companies. For example, by logging in these databases it is possible, at the moment of the quote request, to retrieve useful insurance specific information such as the number of claims caused by the customer in the previous years or how long he has been covered and useful information on his vehicle such as when it has been registered or how many changes of ownership did it experienced.

One of the roles of the actuary is to understand how reliable the information on the policyholder is and to decide how to use that information.

1.3.2 Pricing variables encoding

Formally the pricing variables can be encoded as a vector of real numbers. $\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{ip}) \in \mathcal{X} \subseteq \mathbb{R}^p$. In the modeling framework they can be also called explanatory variables, covariates, predictors or features.

The pricing variables can be of two types:

1. *Quantitative variables*: variables, like policyholder age or vehicle mass, that can be easily represented as a number;
2. *Qualitative variables*: variables, like policyholder sex or vehicle make, that represent a category and are usually represented with strings.

The quantitative variables, with eventual transformations, are already suitable to be used.

To facilitate the use of the qualitative variables, they are usually encoded as sets of binary variables.

If a variable x has only 2 possible modalities, it can be easily encoded in a binary variable z that assigns 0 to one modality and 1 to the other. For example, if $x = \text{sex}$, it can be encoded this way:

$$z = \begin{cases} 1 & \text{if sex} = \text{'Male'} \\ 0 & \text{if sex} = \text{'Female'} \end{cases}$$

In general, if a variable x has K modalities, it can be encoded in $K - 1$ binary variables z_1, z_2, \dots, z_{K-1} . For example, if $x = \text{make}$ and it can have 4 possible modalities ('Fiat', 'Alfa-Romeo', 'Lancia', 'Ferrari') it can be encoded this way:

$$z_1 = \begin{cases} 1 & \text{if make} = \text{'Fiat'} \\ 0 & \text{otherwise} \end{cases}$$

$$z_2 = \begin{cases} 1 & \text{if make} = \text{'Alfa-Romeo'} \\ 0 & \text{otherwise} \end{cases}$$

$$z_3 = \begin{cases} 1 & \text{if make} = \text{'Lancia'} \\ 0 & \text{otherwise} \end{cases}$$

The variables z_1, z_2, z_3 are called dummy variables. We can observe that all the information about the make is embedded in just these 3 variables, so a fourth dummy variable that indicate the modality 'Ferrari' is not needed. Indeed:

$$\text{make} = \text{'Ferrari'} \iff z_1 = z_2 = z_3 = 0$$

In table 1.1 the dummy variable encoding is illustrated.

Table 1.1: Dummy variables encoding.

Make	z_1	z_2	z_3
Fiat	1	0	0
Alfa-Romeo	0	1	0
Lancia	0	0	1
Ferrari	0	0	0

For some models it is suggested to use also the dummy variable that indicates the K^{th} modality. This encoding is called one-hot encoding and it is mainly used in Neural Networks. For the models considered in this paper it is preferred the $K - 1$ dummy variables encoding, so we will always consider it.

In the following, when I use the notation $\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{ip})$, I'll always consider that the qualitative variables have been already encoded as dummy variables, so $(x_{i1}, x_{i2}, \dots, x_{ip}) \in \mathcal{X} \subseteq \mathbb{R}^p$

1.3.3 Pricing Rule and Modeling

The pricing variables are used as input of a *Pricing Rule*.

Definition 1.5 (Pricing Rule). A *Pricing Rule* is a function $f(\cdot)$ that from an instance of a set of pricing variables $\mathbf{x}_i \in \mathcal{X}$ returns a price:

$$\begin{aligned} f: \mathcal{X} &\longrightarrow R_+ \\ \mathbf{x}_i &\longmapsto P_i \end{aligned}$$

The process of pricing consists in defining a Pricing Rule based on observed data from the past and assumptions on the future.

The first step for defining a Pricing Rule is to model the total cost of claims S and obtain a pricing rule for the risk premium $P^{(risk)}$.

Definition 1.6 (Modeling). Modeling a *response variable* Y means finding a function

$$r: \mathcal{X} \rightarrow \mathcal{C}$$

that, given a set of explanatory variables $\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{ip}) \in \mathcal{X} \subseteq \mathbb{R}^p$, returns the expected value of the response variable $E(Y)$ and eventually other moments of Y or even the full distribution of Y .

In definition 1.6 I used a generic \mathcal{C} as codomain of the function $r(\cdot)$ to not specify whether the model describes just $E(Y)$ (and so $\mathcal{C} = \mathbb{R}$) or something more, such as the couple $(E(Y), Var(Y))$ or the full distribution of Y .

As we observed in section 1.2.2, under the compound distribution hypotheses, it is not needed to model directly the total cost of claims S , but we can separately model N and Z .

1.3.4 Response variables and distributions

Usually in statistical modeling, the response variables are seen as random variables with a distribution belonging to a specified family.

1.3.4.1 Distribution for the claims count N

The claim count N is a discrete variable with determination in $\{0, 1, 2, 3, \dots\}$. Even if in practice the number of claims can't be arbitrarily high, N is usually modeled with distributions that give a positive probability to all the numbers in its support. One of the most common distribution used for N is the Poisson distribution.

Definition 1.7 (Poisson Distribution). A random variable N with support $\{0, 1, 2, 3, \dots\}$ has a Poisson distribution, if its probability function is:

$$p_N(n) = P(N = n) = e^{-\lambda} \frac{\lambda^n}{n!}, \quad \lambda > 0$$

We will indicate it with the notation $N \sim \text{Poisson}(\lambda)$.

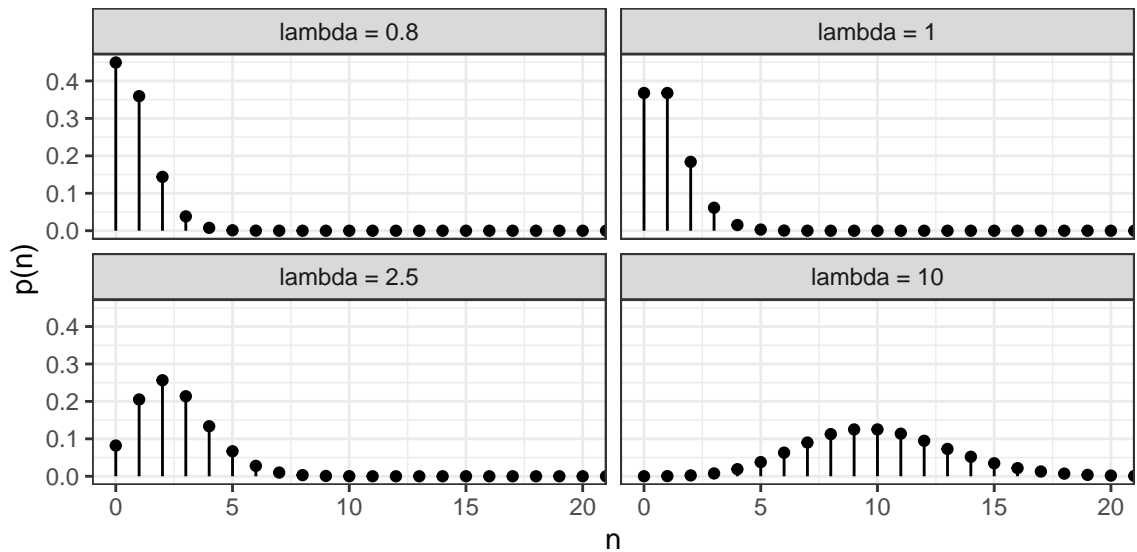


Figure 1.2: Poisson distribution for some values of λ .

The Poisson distribution is a parametric distribution that depends on only the parameter λ . In figure 1.2, for different levels of λ the distribution is represented. These plots show how for larger values of λ , the distribution is shifted to larger values and it is wider.

Indeed, the first two moments are:

$$E(N) = \lambda$$

$$Var(N) = \lambda$$

Thus, increasing λ , both $E(N)$ and $Var(N)$ increase.

Looking to the distribution shape, we can see that:

- if $\lambda < 1$, the mode is in $n = 0$;
- if $\lambda = 1$, $p(0) = p(1) = \frac{1}{e}$;
- if $\lambda > 1$, the mode is in a value greater than 0 and, as λ increases, the distribution assumes a bell shape similar to the Normal distribution one. The convergence to the Normal distribution can be obtained with the *Central Limit Theorem*.

In non-life insurance we usually are in the case with $\lambda < 1$. E.g. the average number of claims for motor third party liability insurances in Italy, in 2018 has been 5.68%¹.

The property $Var(N) = E(N)$ is an important constraint when the distribution is used in practice. It is possible that the observed data shows a different pattern. Often the observed data shows a situation where $Var(N) > E(N)$. This phenomenon is called *overdispersion*.

To address this issue it is possible to use more flexible distributions, such as Negative-Binomial distribution, or to adopt less assumptions on the response variable distribution. One common technique is the assumption of Quasi-Poisson distribution, that we will describe in chapter 2.

1.3.4.2 Exposure

In section 1.1 we said that non-life insurances usually have a fixed coverage period that usually spans for one year. Often we work with portfolios of insurances with different coverage periods. For example, this could be due to the presence of

¹ANIA yearly statistical report for motor third party liability

insurances born with shorter coverage periods or to the presence of insurances that has been closed earlier. Moreover, in companies data, often insurance data are collected for accounting years. This means that, if an insurance coverage c spans in two consecutive years a and $a + 1$, it is collected as two records: the couple (c, a) and the couple $(c, a + 1)$. This situation is quite common, as usually coverages start during the year and not all at the first of the year.

The coverage span for an insurance coverage is called *exposure* and it is usually measured in years-at-risk. For instance, if an insurance coverage spans for 3 months, it corresponds to a quarter of year, so the exposure, measured in years-at-risk, is $v = \frac{1}{4}$. The term year-at-risk comes from the fact that the policyholder exposure is a risk for the insurer, so the exposure is the period in which the insurer is exposed to the risk of paying claims.

It is natural to assume that, if a policyholder has a longer exposure, it is expected for him to experience more claims. Considering that we have to work with policies with different exposures, in order to take this aspect into account, the usual assumption taken in the following. Said M the number of claims the policyholder will experience during his period of exposure v and N the number of claims the policyholder would experience during one year, we assume $E(M) = vE(N)$.

This assumption can be further extended if we assume that the claims come from a *Poisson process*.

Definition 1.8 (Counting Process). A stochastic process $\{N(t), t \geq 0\}$ is called *counting process* if:

1. The determination of $N(t)$ are natural numbers

$$N(t) \in \{0, 1, 2, \dots\} \quad t \geq 0$$

2. The process is not decreasing

$$s < t \Rightarrow N(s) \leq N(t)$$

In a counting process $\{N(t), t \geq 0\}$:

- $N(t)$ can be interpreted as the number of events or arrivals that occur in the period $[0, t]$;

- $N(t) - N(s)$, $s \leq t$ can be interpreted as the number of events or arrivals that occur in the period $]s, t]$. $N(t) - N(s)$ is also called *increment* of the process.

The counting process can be used to model the number of claims that occur to a specific policy.

Definition 1.9 (Poisson Process). A counting process $\{N(t), t \geq 0\}$ is a *Poisson process* with intensity λ if:

1. The increments of the process are stochastically independent
 $\forall n \geq 0, \forall s_1 < t_1 \leq \dots \leq s_n < t_n$
 $\Rightarrow N(t_1) - N(s_1), \dots, N(t_n) - N(s_n)$ are stochastically independent;
2. The probability of arrival in an interval is proportional to the size of the interval
 $\forall t \geq 0, \forall \Delta t > 0 \Rightarrow P(N(t + \Delta t) - N(t) = 1) = \lambda \Delta t + o(\Delta t)$
 where $\lim_{\Delta t \rightarrow 0} \frac{o(\Delta t)}{\Delta t} = 0$
3. Multiple arrivals are excluded
 $\forall t \geq 0, \forall \Delta t > 0 \Rightarrow P(N(t + \Delta t) - N(t) \geq 2) = o(\Delta t)$
4. Arrivals at time 0 are almost impossible
 $P(N(0) = 0) = 1$

Under these hypotheses we obtain the following result:

Theorem 1.1 (Poisson Process). *If $\{N(t), t \geq 0\}$ is a Poisson process with intensity λ , then:*

$$\forall t \geq 0, \forall \Delta t > 0, \Rightarrow N(t + \Delta t) - N(t) \sim \text{Poisson}(\lambda \Delta t)$$

This result tells us that the distribution of the number of events in any interval $]t, t + \Delta t]$ only depends on the size of the interval Δt . Moreover, for the Poisson property we saw in section 1.3.4.1, we get:

$$E(N(t + \Delta t) - N(t)) = \lambda \Delta t$$

So, the expected number of arrivals is proportional to the size of the interval Δt . The intensity of the process λ can be also interpreted as the expected number of claims in a unitary period.

If we assume that the claims that occur to a policy comes from a Poisson process with intensity λ , if we observe that policy for the period $]t, t + v]$, the claims count in that exposure period M are distributed as:

$$M \sim \text{Poisson}(\lambda \Delta t)$$

In particular, if the observed period spans 1 year, we get:

$$M = N \sim \text{Poisson}(\lambda)$$

1.3.4.3 Distribution for the claim severity Z

The claim severity Z is a continuous variable with determination in $[0, +\infty[$. As for the claims count N , even if in practice it can't be arbitrarily high, it is usually modeled with distributions that give a positive probability to all the numbers in $]0, +\infty[$. As the null claims are excluded, it is natural to assume $P(Z = 0) = 0$. One of the most common distribution used for Z is the Gamma distribution.

Definition 1.10 (Gamma Distribution). A random variable Z with support $[0, +\infty[$ has a Gamma distribution, if its probability density function is:

$$f_Z(z) = \frac{\rho^\alpha}{\Gamma(\alpha)} z^{\alpha-1} e^{-\rho z}, \quad \alpha > 0, \rho > 0$$

where $\Gamma(\alpha) = \int_0^{+\infty} z^{\alpha-1} e^{-z} dz$.

We will indicate it with the notation $Z \sim \text{Gamma}(\alpha, \rho)$.

The Gamma distribution is a parametric distribution that depends on two parameters:

- $\alpha > 0$, called shape parameter
- $\rho > 0$, called scale parameter

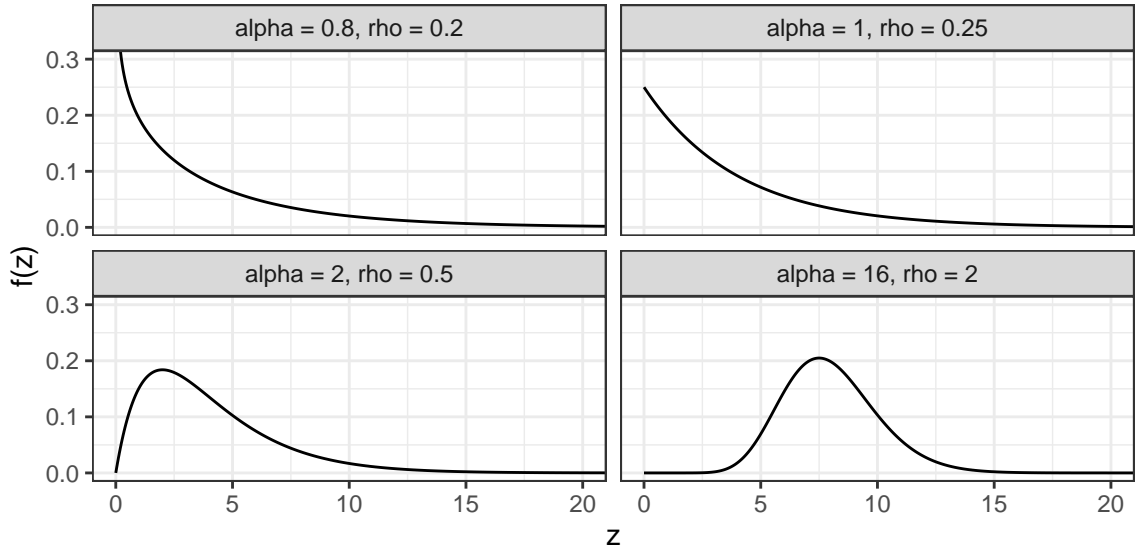


Figure 1.3: Gamma distribution for some values of α and ρ .

The first two moments of the Gamma distribution are:

$$E(Z) = \frac{\alpha}{\rho}$$

$$Var(Z) = \frac{\alpha}{\rho^2}$$

In figure 1.3, for different levels of α and γ the distribution is represented. These plots show how changing the values of α and γ , the shape changes. We can see that:

- if $\alpha < 1$, $f_z(\cdot)$ is not defined in 0 and it has a vertical asymptote in $z = 0$. In $]0, +\infty]$ it is monotonically decreasing.
- if $\alpha = 1$, $f_z(\cdot)$ starts from $f(0) = \rho$ and then decreases monotonically. In this case, the density function becomes $f_z(z) = \rho e^{-\rho z}$ and the distribution is also called exponential distribution.
- if $\alpha > 0$, $f_z(\cdot)$ starts from $f(0) = 0$, increases until the mode and then decreases.

In figure 1.3 the first three distributions represented have the same expected value $E(Z) = \frac{\alpha}{\rho} = 4$, but different shapes. The third and the fourth have the

same variance $Var(Z) = \frac{\alpha}{\rho^2} = 8$, but different expected values. As the shape parameter α increases, the distribution assumes a bell shape similar to the Normal distribution one. The convergence to the Normal distribution can be obtained with the *Central Limit Theorem*.

Another parametrization often used for Gamma distribution is obtained by using the mean μ as a parameter:

$$\mu = \frac{\alpha}{\rho}$$

With this parametrization, the density function becomes:

$$f_Z(z) = \frac{\left(\frac{\alpha}{\mu}\right)^\alpha}{\Gamma(\alpha)} z^{\alpha-1} e^{-\frac{\alpha}{\mu}z}, \quad \alpha > 0, \rho > 0$$

The advantage of using the parameters (α, μ) is that the link between $E(Z)$ and $Var(Z)$ becomes clearer:

$$\begin{aligned} E(Z) &= \mu \\ Var(Z) &= \frac{1}{\alpha} \mu^2 \end{aligned}$$

Computing the coefficient of variation we then obtain:

$$CV(Z) = \frac{\sqrt{Var(Z)}}{E(Z)} = \frac{1}{\sqrt{\alpha}}$$

This result means that, given the shape parameter α , the coefficient of variation is constant. As we saw for Poisson distribution, it is possible that observed data shows a different pattern. In chapter 2, for Gamma distribution, we will use the parametrization based on (α, μ) instead of the one based on (α, ρ) .

Another characteristic of Gamma distribution that could be problematic in modeling claims size is that it has a light tail. This means that, as z goes to $+\infty$,

$f_Z(z)$ approaches 0 quite fast. This could lead to a poor fitting for *large claims*. Other distributions with havier tails are for example the *log-Normal* and the *Pareto*.

1.3.4.4 Large claims

Modeling large claims in quite difficult in practice because usually there is not a lot of observed data on them, so it is hard to understand if they are related to some risk factors (identifiable by the pricing variables) or they happen just by chance.

First of all, to model large claims, we must define what a large claim is. What is usually done in practice is just choosing a threshold \bar{z} and considering large all the claims with a size that exceeds that threshold. The value \bar{z} must be chosen sufficiently big to consider large the claims above \bar{z} , but not so big that there are not enough observed claims that exceeds \bar{z} . One common choice for Motor Third Party Liability in European markets could be $\bar{z} = 100'000\text{€}$.

Definition 1.11 (Large and Attritional Claims). Given a predetermined threshold \bar{z} , we say that:

- a claim Z is a *large claim* if $Z > \bar{z}$
- a claim Z is an *attritional claim* if $Z \leq \bar{z}$

For each claim Z we call:

- *Capped Claim Size*
 $Z' = \min(Z, \bar{z});$
- *Excess Over the Threshold*
 $Z'' = \max(Z - \bar{z}, 0).$

In figure 1.4 the *Capped Claim Size* and the *Excess Over the Threshold* are shown. It is easy to show that Z can be decomposed as:

$$Z = Z' + Z''$$

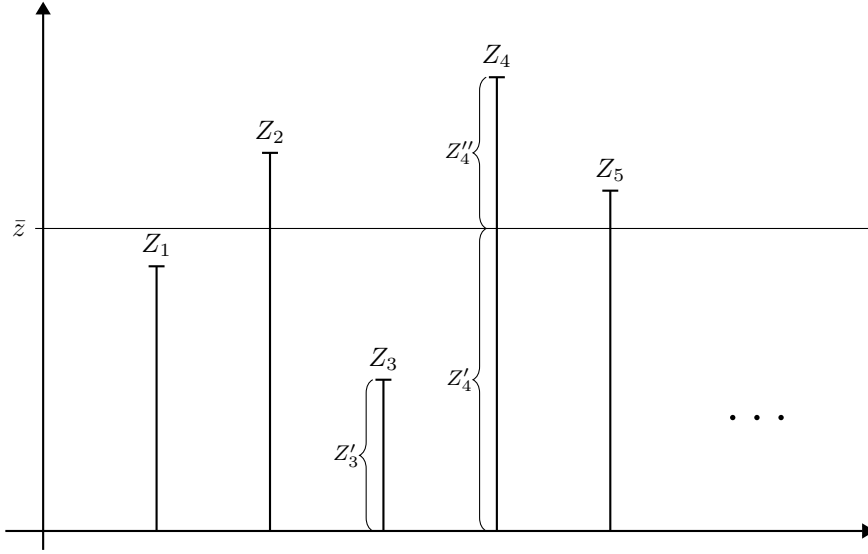


Figure 1.4: Large claims.

Given the total number of claims N , it can be decomposed as:

$$N = N^{(a)} + N^{(l)}$$

where

- $N^{(a)}$ is the attritional claims count, i.e. the number of claims with size $Z \leq \bar{z}$;
- $N^{(l)}$ is the large claims count, i.e. the number of claims with size $Z > \bar{z}$;

Let's indicate with $Z_{(i)}$ the i^{th} in order from the smallest to the bigger. Sorting the claims we can separate the attritional claims from the large claims as follows:

$$\underbrace{Z_{(1)}, Z_{(2)}, \dots, Z_{(N^{(a)})}}_{\text{Attritional Claims}}, \underbrace{Z_{(N^{(a)}+1)}, Z_{(N^{(a)}+2)}, \dots, Z_{(N^{(a)}+N^{(l)})}}_{\text{Large Claims}}$$

In order to model the large claims it is possible to use the following three decompositions of the total cost of claims S :

$$\begin{aligned}
 S &= \underbrace{Z_{(1)} + Z_{(2)} + \dots + Z_{(N^{(a)})}}_{\text{Attritional Claims}} + \underbrace{Z_{(N^{(a)}+1)} + Z_{(N^{(a)}+2)} + \dots + Z_{(N^{(a)}+N^{(l)})}}_{\text{Large Claims}} \\
 &= \underbrace{\sum_{i=1}^{N^{(a)}} Z_{(i)}}_{=S^{(a)}} + \underbrace{\sum_{i=N^{(a)}+1}^{N^{(a)}+N^{(l)}} Z_{(i)}}_{=S^{(l)}} = S^{(a)} + S^{(l)}
 \end{aligned} \tag{1.11}$$

$$S = \sum_{i=1}^N Z_i = \sum_{i=1}^N (Z_i I_{Z_i > \bar{z}} + Z_i I_{Z_i \leq \bar{z}}) \tag{1.12}$$

$$S = \sum_{i=1}^N Z_i = \sum_{i=1}^N (Z'_i + Z''_i) = \sum_{i=1}^N (Z'_i + Z''_i I_{Z_i > \bar{z}}) \tag{1.13}$$

These three decompositions of S are useful because they provide three decompositions of $E(S)$:

$$\begin{aligned}
 E(S) &= E(S^{(a)}) + E(S^{(l)}) \\
 &= E(N^{(a)})E(Z|Z \leq \bar{z}) + E(N^{(l)})E(Z|Z > \bar{z})
 \end{aligned} \tag{1.14}$$

$$\begin{aligned}
 E(S) &= E(N)E(Z) \\
 &= E(N) [P(Z \leq \bar{z})E(Z|Z \leq \bar{z}) + P(Z > \bar{z})E(Z|Z > \bar{z})] \\
 &= E(N) [(1 - P(Z > \bar{z}))E(Z|Z \leq \bar{z}) + P(Z > \bar{z})E(Z|Z > \bar{z})]
 \end{aligned} \tag{1.15}$$

$$\begin{aligned}
 E(S) &= E(N)E(Z) \\
 &= E(N) [E(Z') + P(Z > \bar{z})E(Z'')]
 \end{aligned} \tag{1.16}$$

1.14, 1.15 and 1.16 provide three approaches to model attritional and large claims.

1. Looking to 1.14 we can model separately attritional claims and large claims. Modeling $N^{(a)}$ and $Z|Z \leq \bar{z}$ we estimate the total cost of claims for the

attritional part $S^{(a)}$; modeling $N^{(l)}$ and $Z|Z > \bar{z}$ we estimate the total cost of claims for the large part $S^{(l)}$.

2. Looking to 1.15 we can model together the claim count N , and then we can model the cost of the attritional claims $Z|Z \leq \bar{z}$, the cost of the large claims $Z|Z > \bar{z}$ and the probability to exceed the threshold $P(Z > \bar{z})$.
3. Looking to 1.16 we can model together the claim count N , and then we can model the capped claims size Z' , the excess over the threshold Z'' and the probability to exceed the threshold $P(Z > \bar{z})$.

If the large claims component weights a lot on the total cost of claims, these approaches could lead to quite different estimations of $E(S)$. In particular, if in the observed data the number of large claims is small, it will be hard to model both $N^{(l)}$ and $P(Z > \bar{z})$, so the modeling process could lead to a flat or almost flat model for these components. However, with the first approach, a flat model for $N^{(l)}$ leads to distribute the observed total cost of large claims proportionally to all the policies, while with the second and the third, a flat model for $P(Z > \bar{z})$ leads to distribute the observed total cost of large claims proportionally to the expected number of claims $E(N)$. So, with the first approach, a flat model brings to more solidarity between policies, while, with the second approach, a flat model could bring to an exacerbation of the differences identified by modeling N .

For the second approach we must also introduce a distribution suitable for modeling $P(Z > \bar{z})$.

1.3.4.5 Binomial distribution

The *binomial distribution* is used to model the counting on events that occurs (successes) in a fixed amount of trials n . For example we can use it to model the number of large claims withing a fixed number of n claims.

Definition 1.12 (Binomial Distribution). A random variable Y with support $\{0, 1, 2, \dots, n\}$ has a Binomial distribution, if its probability function is:

$$p_Y(y) = P(Y = y) = \binom{n}{y} p^y (1 - p)^{n-y}, \quad p \in [0, 1]$$

We will indicate it with the notation $Y \sim \text{Binom}(n, p)$.

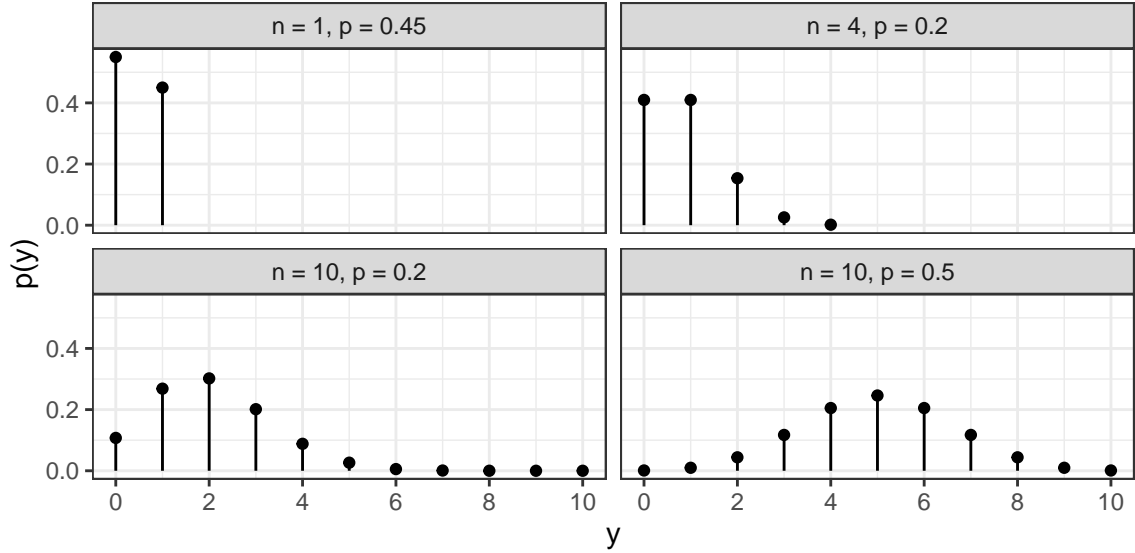


Figure 1.5: Binomial distribution for some values of n and p .

The binomial distribution is a parametric distribution that depends on the parameters n and p . n represents the number of trials, while p represents the probability for a trial to success. The assumption is that the n trials are identical, so they have all the same probability p to success. In figure 1.5, for different levels of n and p , the distribution is represented.

The first two moments of the binomial distribution are:

$$E(N) = np$$

$$\text{Var}(N) = np(1 - p)$$

If $n = 1$, the binomial distribution assumes only the values 1 (with probability p) and 0 (with probability $1 - p$). In this case it is also called *Bernoullian distribution* and it can be used to model the indicator of an event I_E .

If $n > 1$, the binomial distribution assumes a shape centered on its expected value $E(Y) = np$ and fading for values of y that moves away from $E(Y)$. As n

increases, the distribution assumes a bell shape similar to the Normal distribution one. The convergence to the Normal distribution can be obtained with the *Central Limit Theorem*.

From the binomial Distribution it is also possible to define the scaled binomial distribution by dividing its value by n .

Definition 1.13 (Scaled Binomial Distribution). If $Y \sim \text{Binom}(n, p)$, and $Y' = \frac{Y}{n}$, we will say that Y' has a *Scaled Binomial Distribution* and we will indicate it with the notation $Y' \sim \text{Binom}(n, p)/n$.

The support of Y' is $\{0, \frac{1}{n}, \frac{2}{n}, \dots, 1\}$ and its probability function is:

$$p_{Y'}(y') = P(Y' = y') = \binom{n}{ny'} p^{ny'} (1-p)^{n-ny'}, \quad p \in [0, 1]$$

In chapter 2 we will use the Scaled Binomial Distribution.

In non-life insurance pricing, the binomial distribution can be used to model the probability for a claim to have specific characteristics. For example we can use it to model the probability that a certain claim is a large one $P(Z > \bar{z})$ in order to model separately attritional claims severity $\{Z|Z \leq \bar{z}\}$ and large claims severity $\{Z|Z > \bar{z}\}$, as we have seen in section section 1.3.4.4.

Another example is the decomposition between claims with only material damages and claims with also bodily injuries. Modeling separately these two components is useful because they usually have a different distribution for the claim size.

As for large claims we can decompose S in the following two ways:

$$\begin{aligned} E(S) &= E(S^{(things)}) + E(S^{(inj)}) \\ &= E(N^{(things)})E(Z|\bar{J}) + E(N^{(inj)})E(Z|J) \end{aligned} \quad (1.17)$$

$$\begin{aligned} E(S) &= E(N)E(Z) \\ &= E(N) [P(\bar{J})E(Z|\bar{J}) + P(J)E(Z|J)] \\ &= E(N) [(1 - P(J))E(Z|\bar{J}) + P(J)E(Z|J)] \end{aligned} \quad (1.18)$$

where:

- $N^{(things)}$ is the number of claims with only material damages;
- $N^{(inj)}$ is the number of claims with injuries;
- J is the event that represent that a specific claim presents injuries; such as Z is a representative for Z_1, Z_2, \dots, Z_N , J is a representative for J_1, J_2, \dots, J_N .

Combining this decomposition with what we have seen in large claims decomposition, we can further develop our decomposition taking into account both the presence or absence of injuries and the occurrence or not of a large claim. One example could be:

$$\begin{aligned}
 E(S) &= E(N) [(1 - P(J)) E(Z|\bar{J}) + P(J) E(Z|J)] \\
 &= E(N) \{ \\
 &\quad (1 - P(J)) E(Z|\bar{J}) \\
 &\quad + P(J) [\\
 &\quad \quad P(Z \leq \bar{z}|J) E(Z | Z \leq \bar{z} \wedge J) \\
 &\quad \quad + P(Z < \bar{z}|J) E(Z | Z < \bar{z} \wedge J) \\
 &\quad] \\
 &\quad \}
 \end{aligned}$$

This way, we are decomposing only the claims with injuries between attritional and large. That makes sense because claims that don't produce injuries usually have small severities.

1.3.5 Model fitting and data available

Once we have chosen how to decompose S , we have to model the response variables needed for that decomposition (N, Z, I_J, \dots) with the explanatory variables. Thus we have to estimate a function $r : \mathcal{X} \rightarrow \mathcal{C}$ as defined in 1.6.

In order to estimate $r(\cdot)$ we have also to take some assumptions on the distribution of the response variable and on the shape of $r(\cdot)$. We will call *model*

a set of assumptions on the response variable and on the shape of $r(\cdot)$. We will discuss some of the most widespread models for claims count and claims severity in chapter 2.

Defined the model, we have to estimate it using observed data. In general, to model a response variable Y_i with the explanatory variables $\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{ip}) \in \mathcal{X} \subseteq \mathbb{R}^p$, the observed data is in the form:

$$\mathcal{D} = \{(\mathbf{x}_1, w_1, y_1), (\mathbf{x}_2, w_2, y_2), \dots, (\mathbf{x}_i, w_i, y_i), \dots, (\mathbf{x}_n, w_n, y_n)\}$$

where:

- n is the number of observations in the dataset;
- $\mathbf{x}_i \in \mathcal{X} \subseteq \mathbb{R}^p$ is the set of explanatory variable for the observation i ;
- w_i is the weight for the observation i ;
- $y_i \in \mathcal{Y} \subseteq \mathbb{R}$ is the realization of the response variable Y_i for the observation i .

What an observation is, depends on the variable we are modeling. For instance:

- If we are modeling the yearly claim count N_i , each observation could be a policy (or a couple (policy, accounting year)), the weights could be the exposures v_i and the realizations of response variables could be the number of observed claims for that policy (or couple (policy, accounting year)).
- If we are modeling the claim severity Z_j , each observation could be a claim j , the weights could all be 1 and the realizations of response variables could be the observed cost for the claim j . It is also possible to model the claim severity taking into account the total cost of claims for the policy $S_i = \sum_{j=1}^{N_i} Z_j$. In this case, each observation would be a policy i , the weights would be the number of claims for each policy n_i and the realizations of response variables would be the total observed cost for the claims of the policy i .
- If we are modeling the occurrence of injuries in a claim I_{Jj} , each observation could be a claim j , the weights could be all 1 and the realizations of response variables could be an indicator that assume the value 1 if the claim j caused

injuries and 0 otherwise. As for the claim severity, we can also aggregate data for policy, so each observation would be a policy i , the weights would be the number of claims n_i for the policy i and the realizations of response variables would be the number of claims that caused injuries among the claims of the policy i .

In each of these cases, y_i is seen as a realization of the random variable Y_i . With an inferential process we obtain estimations on Y_i distribution based on observations of their realizations y_i .

1.3.5.1 Settlement process and IBNR

One of the challenges in non-life insurance pricing is that obtaining the observed data is not so straightforward. In many insurance coverages, such as MTPL, the settlement process could last many years, so, if we want to develop models using data from recent years, not all the information is available. To better understand this aspect we have to discuss how the settlement process works.

In figure 1.6 the settlement process for a claim is represented. At time t_1 the insured event (e.g. an accident) occurs. From this moment a liability for the insurer emerges, even if the insurer has not been notified yet. This liability is called *Outstanding Loss Liability*. In t_2 the claim is reported and the insurance is notified about the occurrence of the event. From this moment the settlement process starts. This process consists in evaluating the event and understanding the responsibilities of the parts and the entity of the damage. During this process, controversies between the parts can emerge and, in particular if injuries occurred, the damage evaluation can take a lot of time. When the situation is clear and everything is defined, the claim is settled and the liabilities are paid. In t_3 we have the settlement and in t_4 the claim is closed. It is possible that $t_4 = t_3$, but in general it is not always the case. If the settlement process takes a long time and the insurer already knows he will have to pay something, he can pay some partial payments during the period $[t_2, t_3]$. These intermediate payments are paid at times $\tau_1, \tau_2, \dots, \tau_n \in [t_2, t_3]$. It is also possible that a claim is opened and then gets closed without any payment. After the closing (t_4) it is also possible that a claim is reopened and that more payments emerge.

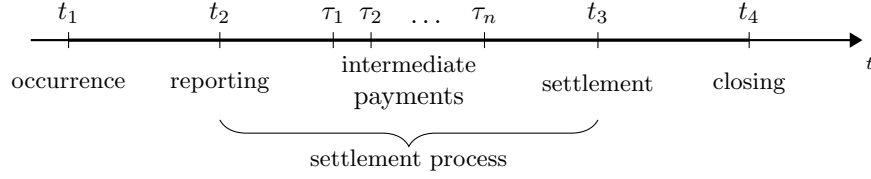


Figure 1.6: Claim timeline.

From the moment the claim is reported (t_2), the insurer estimates how much he is going to pay for that claim and he allocates that sum in a reserve. As new information emerges and some payments are settled, the reserve is updated. The aim for this reserve is to have a best estimate for the future payments for the claims already emerged. As the claim gets settled, the sum between the payed and the reserved converges to the final cost of the claim.

From this description emerges that:

- In the period $]t_1, t_2[$ the insurer has an outstanding loss liability for an event that has not been reported yet; in this case we will talk about *Incurred But Not Yet Reported* claim (IBNyR).
- In the period $[t_2, t_3[$ the insurer has an outstanding loss liability for an event that has been reported, but has not been totally settled yet, so that liability is just an estimate; in this case we will talk about *Incurred But Not Enough Reported* claim (IBNeR).

1.3.5.2 Model fitting with available data

The IBNyR and IBNeR issue is particularly challenging when we have to perform a risk evaluation at a specific time t . In general t_1, t_2, \dots are not known a priori, so we don't know if in the future more claims for accidents occurred in the past will be reported and we don't know if the ones that are already reported will experience a revaluation. That means that, in general, when we model N and Z at a specific time t , we can't observe the total number of claims occurred to each policy n_i and the payments for each claim z_j . What we can use is:

- $n_i^{(t)} = n_i^{(\text{reported in } t)}$
where:

- $n_i^{(\text{reported in } t)}$ is the number of reported claims in t for the policy i ;
- $z_j^{(t)} = z_j^{(\text{payed in } t)} + z_j^{(\text{reserved in } t)}$
 where:
 - $z_j^{(\text{payed in } t)}$ is the amount already payed in t for the claim j ;
 - $z_j^{(\text{reserved in } t)}$ is the amount reserved in t for the claim j .

When we use this data for modeling the total cost of claims we must be particularly aware on what we are using. In general:

$$\begin{aligned} n_i^{(t)} &\neq n_i \\ z_j^{(t)} &\neq z_j \end{aligned}$$

The common case is that $n_i^{(t)} < n_i$ and $z_j^{(t)} < z_j$. If we used $n_i^{(t)}$ and $z_j^{(t)}$ without any correction, we would underestimate both $E(N)$ and $E(Z)$, obtaining a distorted estimate for $E(S)$.

To tackle these problems what is usually done is fitting the models for S_i with $n_i^{(t)}$ and $z_j^{(t)}$ and then apply a flat corrective coefficients α to $\widehat{E}(S_i)$ based on an aggregated estimate of $E(S)$ that takes into account the long settlement process.

An estimate for the expected total cost of claims for a generic policy in the portfolio $E(S)$ can be obtained with techniques based on runoff triangles, such as the *Chain Ladder*. These techniques are based on projecting the cost of claims already emerged to the final total cost of claims. We are not going to discuss these techniques in this thesis. We just have to know that these techniques provide us with an estimate for $E(S)$. Let's call it $\widehat{E}(S)^{CL}$. This estimate does not depend on explanatory variables; it is a sort of average total cost of claims for the policies in the portfolio.

Meanwhile, with the available data $n_i^{(t)}$ and $z_j^{(t)}$, the fitting for all the models needed in the decomposition of S is performed and, for each policy $i \in \{1, 2, \dots, n\}$ for $E(S_i)$ is obtained. Let's call it $\widehat{E}(S_i)'$. As we used the data available in t that comes from claims not totally settled, $\widehat{E}(S_i)'$ is a distorted estimate for $E(S_i)$.

We can then balance the estimates $\widehat{E(S_i)}$ with $\widehat{E(S)}^{CL}$ by computing:

$$\alpha = \frac{n}{\sum_{i=1}^n \widehat{E(S_i)}} \widehat{E(S)}^{CL}$$

and applying to the estimates as follows:

$$\widehat{E(S_i)} = \alpha \widehat{E(S_i)}'$$

We will call $\widehat{E(S_i)}$ rebalanced estimates.

The property of these balanced estimates is that on average they point to $\widehat{E(S)}^{CL}$:

$$\begin{aligned} \frac{\sum_{i=1}^n \widehat{E(S_i)}}{n} &= \frac{\sum_{i=1}^n \alpha \widehat{E(S_i)'}}{n} \\ &= \alpha \frac{\sum_{i=1}^n \widehat{E(S_i)'}}{n} \\ &= \frac{n}{\sum_{i=1}^n \widehat{E(S_i)'}} \widehat{E(S)}^{CL} \frac{\sum_{i=1}^n \widehat{E(S_i)'}}{n} \\ &= \widehat{E(S)}^{CL} \end{aligned}$$

So, if $\widehat{E(S)}^{CL}$ is a unbiased estimator for $E(S)$, we obtain:

$$E\left(\frac{\sum_{i=1}^n \widehat{E(S_i)}}{n}\right) = E\left(\widehat{E(S)}^{CL}\right) = E(S)$$

This procedure can be further developed by balancing not directly the total cost of claims $E(S)$, but its components. For example, we could separately balance the total cost of claims that only caused damage to things and the total cost of claims that caused injuries. This separation in components can lead to a more precise estimate because usually claims that caused injuries have a slower settlement process so they will have a higher corrective coefficient α .

If the dataset contains policies from many years and during last years a relevant

change in the portfolio risk mixture happened, it is also possible to compute α not with all the n policies of the dataset, but only with the policies from the last year of the dataset.

The fact that the final estimates $\widehat{E}(S_i)$ are rebalanced on $\widehat{E}(S)^{CL}$ means that the explanatory variables effects estimated with $n_i^{(t)}$ and $z_j^{(t)}$ are used just as relative effects and not absolute ones. For instance, if the model says that young people have an expected total cost of claims $\widehat{E}(S_i)'$ that is two times the old people one, that relative coefficient 2 will be kept also in the balanced estimate $\widehat{E}(S_i)$.

For this reason, in practice, often the modeling is considered composed in 2 parts:

1. *Tariff Requirement* (or *Fabbisogno Tariffario*): the estimate of $\widehat{E}(S)^{CL}$ by aggregated data;
2. *Personalization*: the estimate of $\widehat{E}(S_i)'$ and the relative coefficients.

The techniques used for *Tariff Requirement* are employed also to estimating the general liability position for the company. This information is particularly important and it is reported in the company financial statement.

1.4 Beyond technical pricing

In section 1.2.3 we defined:

- the *Risk Premium*

$$P^{(risk)} = E(S)$$
- the *Technical Price*

$$P^{(tech)} = E(S) + E$$

In section 1.3 we described how the risk premium can be estimated. In this thesis we are not going to deal with the estimate of the expenses.

In this section we are going to discuss what the *Tariff* and the *Offer Price* are and which are the further needs that the offer should satisfy. The following description is referred to MTPL insurance in the Italian market. Most of the comments we make can be applied to other motor coverages too.

1.4.1 Tariff and Offer Price

The *Tariff* is the official price for the policy. Over the cost of claims and the expenses, it must include all the loadings for cost of capital and profits. The tariff has a particular importance because it is subjected to strict regulations and it must be approved by the supervisory authority, that in Italy is the IVASS (Istituto per la vigilanza sulle Assicurazioni).

In section 1.3.1 we described some of the explanatory variables that can be used to build the technical price. For technical pricing there are no constraints because it is used only for internal monitoring and the final price proposed to the client does not directly depends on it. However, some of the variables used for technical pricing can't be used in tariff. In particular, the regulations say that companies can't discriminate clients based on sex, ethnic group, religion or place of birth. Thus, for example, even if from statistical data we see that women usually experience less claims than men, we can't discriminate men by offering them a higher price. Moreover, some variables have constraints on tariff coefficients. For example, in MTPL insurance, the bonus-malus class is strongly regulated. Every company must recognize the bonus-malus class matured by clients (even if they matured them with other companies) and the coefficients of this variable must be monotonically increasing, i.e. a lower class must correspond to a better tariff. We remind that in the Italian bonus-malus system the lower the class the better the premium. Another tariff constraint is that for some coverage, such as MTPL, the insurer has an obligation to contract. That means that whoever the client is, independently to how risky he is, the company must offer a premium and, if the client accepts, the company must underwrite the insurance contract. In this context, if the company offers an unreasonably high premium, it could fall in an attempt to avoiding the obligation to contract. For this reason, the tariff can't be arbitrarily high and must contemplate a maximum premium. To be sure that all the constraints has been respected, the tariff, before entering in production, must follow a strict approval process.

To make the offer price more flexible and to facilitate business competition, the supervisory authority allows insurance companies to sell policies not to the

tariff price, but to the price obtained subtracting from it a discount $D_i \geq 0$. The premium obtained this way is called *Offer Price*.

$$P_i^{(\text{offer})} = P_i^{(\text{tariff})} - D_i$$

That means that, for the offer price to adequately cover the cost of claims and expenses, the tariff must include a loading for discounting. This loading for discounting, called *discounting flexibility*, can be partially spent by the agent and partially by the insurance company itself. The discounts can be changed over time in a much more agile way than the tariff. For example in Italy, during the Spring 2020 Covid19 crisis, many companies introduced measures to support clients needs with important discounts on both new business and renewal. From a technical point of view, these discounts have been funded by the remarkable decrease on claim frequency due to the traffic decreasing. Discount measures like these are welcomed by the supervisory authority because they promote business competition and lead to lower prices for consumers.

1.4.2 Commercial price

Both tariff and offer price must be based not only on technical logic, but also on commercial one. They are determined with a process of *Price Optimization*. The final goal for a company is to maximize profits this year and in the next ones, so the objective of commercial pricing must be obtaining the optimal price to reach this goal. Maximizing profits is a quite generic goal and can't be easily expressed as an analytical optimization problem. For this reason the pricing choices can be guided by the business strategy that can be translated in specific *Key Performance Indicators* (KPI) that have to be optimized. In this optimization framework, the technical price can be seen as an estimate of the expected cost related to the policy. Knowing the costs it is possible to tune the final premium by working on margins.

The components that act on price optimization can be addressed to:

1. *technical pricing*;
2. *client expectation*;
3. *business strategy*.

We already extensively covered technical pricing in previous chapters.

Client expectation is basically the price that the client is willing to pay for the specific product. Here the basic idea is that if the client would pay a premium much higher than the technical one, it wouldn't make sense for the company to sell him the product at technical price. This is a big opportunity to gain margins. To analyze client expectation, what is usually done is:

- for new business modeling his conversion probability
- for renewal business modeling his retention probability.

For example some guarantees or some options are perceived by the clients as being really worth even if their technical price is not so high. The perception of the client depends also on the competitors pricing and how easy comparing offers from different companies is. In the last years, in the Italian market, the development of aggregators has made much easier for consumers to compare offers from different companies, increasing the competition and the attention on pricing. Anyway, if a company is able to differentiate itself from the others and to make its product been perceived as more valuable, it can sell it at a higher price than other companies. For example this can be achieved by improving customer service and customer experience.

If the technical price and the conversion probability function are given, finding the optimal price for a policy can be expressed as an analytical optimization problem. However, to find the optimal price, one should also take into account that usually policies are not sold alone, but in packages of guarantees. With a wider vision, a business strategy could be selling MTPL policy with almost no margins if it allows to sell other guarantees with higher margins. Moreover, as the aim is not to be profitable this year, but also in the following ones, the company should also consider the *lifetime value* of the client. Indeed, a satisfied client will also stipulate other contracts in the future and can bring to the company other clients from his connections. So, selling a policy with small margins today can lead to high margins tomorrow in other policies.

The business strategy could also contemplate being more aggressive on certain targets of client and less on others. For example, if the company is particularly

strong in certain regions, it could make sense for it to push in that region to further increase its market share. Vice versa, in regions where the company doesn't sell much, it could be safer not to push too much and to be more careful. In a risk management framework, this can be also interpreted as introducing a further risk margin for clusters where there isn't enough observed data and the lack of information brings to more uncertainty. An aggressive pricing can also make sense for a young company that is growing and it is not supposed to be profitable from the first years. From a marketing point of view this strategy can increase the brand awareness by the clients and can strengthen the company image.

Anyway, a company can't arbitrarily discount policies because an excess in discounting could cause severe drawbacks on a financial perspective. Therefore, an insurance company must always respect the solvency constraints defined by the supervisory authority to safeguard itself from bankruptcy. The company solvency is essential to protect all the stakeholders, that are both the clients and the investors.

1.5 The actuary role

In this technical and commercial pricing framework, the actuary is the one that conduct the analysis and define the pricing rules. The *International Actuarial Association* (IAA) describes the actuaries are “highly qualified professionals who analyze the financial impact of risk for organizations like insurers, pensions fund managers, and more” and it states that their work “requires a combination of strong analytical skills, business knowledge, and understanding of human behavior”.²

First of all, the actuary must master the main statistical and data science techniques used to develop model for technical pricing. On this field, in the last years, the development of machine learning and high performance computing has permitted a huge development of technical pricing allowing actuaries to use much more complex variables and models. However, the actuary does not have just to be an expert in statistics and machine learning. He must be also able to interpret the results he gets with his models and use his expertise to understand if the results he gets are fine for future predictions. As we already mentioned, the pricing rules will

²IAA, [About Actuaries](#)

be used for policies that will be sold in the future, so they have to be defined with a mixture of observation of the past and assumptions on the future. In addition, sometimes it is needed to define price for clusters where the company have no historical data. This can happen when a company is expanding to new customers for example by opening new selling channels or by pushing in regions where its market share is quite small. Furthermore, the lack of historical data can be due to the full sector evolution. For example, in these years, new vehicles, such as electric cars and cars with Advanced Driver-Assistance Systems (ADAS), are spreading. As these vehicles didn't exist in the past, historical data doesn't exist. So, finding the proper pricing is challenging. From the company point of view, positioning with a competitive pricing on these segments is important for future business, but the risk must be properly evaluated. For these kind of tasks, the actuary must have a deep domain knowledge on the field.

In the last years, the increase of competition brought to an increasing in commercial price importance. Now most of the companies build their own conversion and retention probability models and they are developing more complex business strategies. In this context it is fundamental for the actuary to understand the clients behaviors, in order to optimize tariff and offer price. The importance for commercial pricing implicate that the technical pricing must not be conducted independently from commercial pricing. Even in companies where the technical pricing and the commercial pricing are carried out by two separate teams, the two teams have to collaborate and coordinate together. This need have some relevant implications on how technical pricing is conducted that we will further discuss in section [2.3](#).

→

2

Statistical models for Non Life Insurance Pricing

In this chapter we are going to describe some of the most widespread models for technical pricing. For each model we are going to describe its benefits and drawbacks and in section [2.3](#) we will compare them by discussing how they fit the pricing needs.

2.1 Statistical Models

In this section we will start by describing the Generalized Linear Model (GLM), that is the most employed model in technical pricing, to then present some of its advancements: the Elastic Net and the Generalized Additive Model (GAM). After this description we will also present the Gradient Boosting Machine (GBM), that is one of the most effective general purpose machine learning models. This allows us to have a comparison between GLM based models and general purpose machine learning models.

2.1.1 GLM

2.1.1.1 Linear Exponential Family

One of the GLM assumptions is that the response variables belong to a *Linear Exponential Family*. In this section we are going to explain what it is and which distributions fit its definition.

Definition 2.1 (Linear Exponential Family). A Linear Exponential Family \mathcal{F} is a parametrical family of probability distributions with density function (or probability function in the discrete case) that can be expressed in the form:

$$f(y; \theta, \lambda) = \exp \left\{ \frac{y\theta - b(\theta)}{\lambda} \right\} c(y, \lambda), \quad y \in \mathcal{Y} \subseteq \mathbb{R}$$

where:

- $\theta \in \Theta \subseteq \mathbb{R}$ is called *canonical parameter*;
- $\lambda \in \Lambda \subseteq]0, +\infty[$ is called *dispersion parameter*;
- $b : \Theta \rightarrow \mathbb{R}$ is a real function called *cumulant function*;
- $c : (\mathcal{Y}, \Lambda) \rightarrow [0, +\infty[$ is a real function;
- Θ is a non degenerate interval, i.e. $\text{int}\Theta$ is not empty.

An exponential family \mathcal{F} is characterized by the elements $(\Theta, b(\cdot), \Lambda, c(\cdot, \cdot))$. By properly choosing the sets Θ, Λ and the functions $b(\cdot), c(\cdot, \cdot)$, it is possible to obtain many useful families.

It can be easily shown that the families Normal, Poisson, Gamma and Binomial are exponential families. In table 2.1 the characterizations for these exponential families are reported.

The distributions that belong to an exponential family have many useful properties. For example they are provided with all the moments and their moments can be obtained using the derivatives of the cumulative function $b(\cdot)$. If Y is a random variable with distribution belonging to an exponential family \mathcal{F} with parameters θ, λ , its first two moments are:

Table 2.1: Some Linear Exponential Families.

Distribution	Notation	Θ	θ	Λ	λ	$b(\theta)$
Normal	$N(\mu, \sigma^2),$ $\mu \in \mathbb{R}, \sigma > 0$	\mathbb{R}	μ	$]0, +\infty[$	σ^2	$\frac{\theta^2}{2}$
Poisson	$Poisson(\mu),$ $\mu > 0$	\mathbb{R}	$\log(\mu)$	$\{1\}$	1	e^θ
Gamma	$Gamma(\alpha, \mu),$ $\alpha > 0, \mu > 0$	$] -\infty, 0[$	$-\frac{1}{\mu}$	$]0, +\infty[$	$\frac{1}{\alpha}$	$-\log(-\theta)$
Scaled Binomial	$Binom(n, p)/n,$ $n \in \mathbb{N}, p \in]0, 1[$	\mathbb{R}	$\log \frac{p}{1-p}$	$\{\frac{1}{n}\}$	$\frac{1}{n}$	$\log(1 + e^\theta)$

$$E(Y) = b'(\theta) \quad (2.1)$$

$$Var(Y) = \lambda b''(\theta) \quad (2.2)$$

As, within a specified family, the parameters θ and λ determine a distribution, in practical problems the object of estimation will be the couple (θ, λ) . In many problems it is natural to consider distributions from a linear exponential family where the dispersion parameter can be expressed as $\lambda = \frac{\phi}{\omega}$, where $\omega > 0$ is a known *weight* and $\phi > 0$ is a parameter that we will keep calling *dispersion parameter*. In this case, the density of probability function depends on the parameters θ and ϕ and will be expressed as:

$$f(y; \theta, \phi, \omega) = \exp \left\{ \frac{\omega}{\phi} [y\theta - b(\theta)] \right\} c(y, \phi, \omega), \quad y \in \mathcal{Y} \subseteq \mathbb{R}$$

In this case the parameters θ and ϕ will be object of estimation, while ω is an already known value. As we will see later, this representation allows us to consider as known weights:

- the exposure v in the Poisson distribution;
- the number of trials n in the Binomial distribution.

2.1.1.2 Model assumptions

Let's assume that, for n statistical units, the observations $\mathcal{D} = \{(\mathbf{x}_1, \omega_1, y_1), \dots, (\mathbf{x}_n, \omega_n, y_n)\}$ are available, where \mathbf{x}_i is a vector of explanatory variables determinations, ω_i is a known weight and y_i is the response variable determination. $\mathbf{x}_i, \omega_i, y_i$ are all real numbers. The vector $\mathbf{y} = (y_1, \dots, y_n)^t$ is considered a determination of the response random vector $\mathbf{Y} = (Y_1, \dots, Y_n)^t$.

In GLM we assume that:

1. The response variables Y_1, \dots, Y_n are stochastically independent and with probability distribution belonging to a same linear exponential family; i.e. the probability distribution of Y_i has density function (or probability function in the discrete case) that can be expressed as:

$$f(y_i; \theta_i, \phi, \omega_i) = \exp \left\{ \frac{\omega_i}{\phi} [y_i \theta_i - b(\theta_i)] \right\} c(y_i, \phi, \omega_i), \quad y_i \in \mathcal{Y} \subseteq \mathbb{R}$$

We highlight that only θ_i and ω_i depend on i , while the dispersion parameter ϕ is the same for all the observations.

2. The explanatory variables determinations vector $\mathbf{x}_i = (1, x_{i1}, \dots, x_{ip})^t$ affects the probability distribution of the response variable Y_i by the linear predictor:

$$\eta_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip}$$

that is a linear function of the regression parameters $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_p)$.

3. The linear predictor η_i is linked to the expected value of the response variable $\mu_i = E(Y_i)$ by the following relation:

$$g(\mu_i) = \eta_i = \mathbf{x}_i^t \boldsymbol{\beta}$$

where $g : \mathbb{R} \rightarrow \mathbb{R}$ is a monotonic function with continuous first and second derivatives. $g(\cdot)$ is called *link function*.

Often, the assumption 1 is called stochastic assumption, while the 2 and 3 are called structural assumptions.

Let's indicate with \mathbf{X} the design matrix, i.e. the matrix in which each row \mathbf{x}_i represents the vector of the explanatory variables for the observation i and each column \mathbf{x}_j represents the vector of the observations for the explanatory variable j . The design matrix is represented in figure 2.1. The matrix starts with a column of 1s, that is used to model the intercept. Thus, it is a matrix $n \times (p + 1)$. We assume, as it is common in actuarial datasets, that $n > p + 1$.

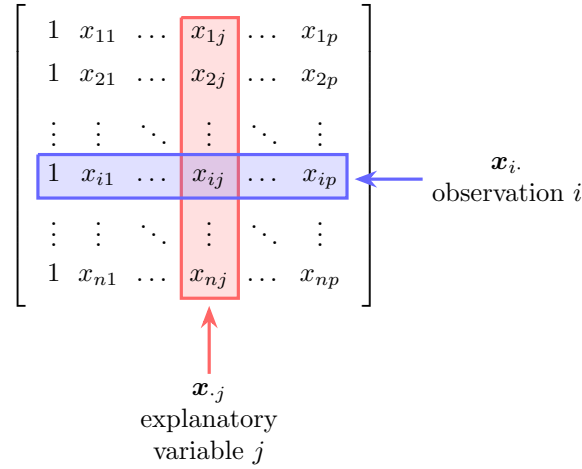


Figure 2.1: Design Matrix \mathbf{X} .

We can then express the GLM structural assumptions in a matrix form as:

$$\mathbf{g}(\boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\beta}$$

where $\mathbf{g}(\cdot)$ must be intended as the vectorial function that links every μ_i to $g(\mu_i)$.

$$\mathbf{g} : \begin{pmatrix} \mathbb{R}^n \\ \mu_1 \\ \vdots \\ \mu_n \end{pmatrix} \longrightarrow \begin{pmatrix} \mathbb{R}^n \\ g(\mu_1) \\ \vdots \\ g(\mu_n) \end{pmatrix}$$

We assume the design matrix to be a full rank matrix, i.e. $\text{rank}(\mathbf{X}) = p + 1$. This assumption corresponds to assuming that the \mathbf{X} columns are linearly independent.

The function $g(\cdot)$ can be chosen as any monotonic function with continuous first and second derivatives. Given a family \mathcal{F} a common choice is its canonical

link function that is defined as:

$$g(\mu) = b'^{-1}(\mu)$$

From (2.1) we we obtains that, as $\mu = b'(\theta)$, choosing the canonical function corresponds to using θ as the linear predictor:

$$\eta = g(\mu) = b'^{-1}(\mu) = \theta$$

In table 2.2 the canonical link functions for the families mentioned in 2.1 are reported.

Table 2.2: Canonical link functions.

Distribution	Cumulant function $b(\theta)$	Derivative $b'(\theta)$	Canonical link function $g(\mu) = b'^{-1}(\mu)$
Normal	$\frac{\theta^2}{2}$	θ	μ
Poisson	e^θ	e^θ	$\log(\mu)$
Gamma	$-\log(-\theta)$	$-\frac{1}{\theta}$	$-\frac{1}{\mu}$
Scaled Binomial	$\log(1 + e^\theta)$	$\frac{e^\theta}{1+e^\theta}$	$\log\left(\frac{p}{1-p}\right)$

In the Gamma case, its canonical function $g(\mu) = -\frac{1}{\mu}$ has the drawback that it links the expected values $\mu \in]0, +\infty[$ to $\eta \in]-\infty, 0[$. This would require some constraints on β because $\eta = \mathbf{x}'\beta$ would have to be < 0 . For this reason, it is preferred to use $g(\mu) = \log(\mu)$ that maps $]0, +\infty[$ to \mathbb{R} .

2.1.1.3 Model fitting

The model depends on the parameters (β, ϕ) . Indeed, the parameters θ_i can be obtained by β as:

$$\theta_i = b'^{-1}(\mu_i) = b'^{-1}(g^{-1}(\eta_i)) = b'^{-1}(g^{-1}(\mathbf{x}_i'\beta))$$

Therefore, fitting the model corresponds to estimating $(\boldsymbol{\beta}, \phi)$. The technique used in GLM is the *maximum likelihood*. Let's indicate with $L(\boldsymbol{\beta}, \phi; \mathbf{y})$ the model likelihood. We remind that the likelihood is a function of the parameters that maps $(\boldsymbol{\beta}, \phi)$ to the density (or probability in the discrete case) of the observed values \mathbf{y} conditioned to the parameters $(\boldsymbol{\beta}, \phi)$

$$\begin{aligned} L: \mathbb{R}^{p+1} \times \Lambda &\longrightarrow [0, +\infty[\\ (\boldsymbol{\beta}, \phi) &\longmapsto f_Y(\mathbf{y}; \boldsymbol{\theta}, \phi) \end{aligned}$$

The maximum likelihood estimates are the values $(\boldsymbol{\beta}, \phi)$ that maximize $L(\boldsymbol{\beta}, \phi; \mathbf{y})$. In practice, $\boldsymbol{\beta}$ are the parameters of interest, while ϕ is considered as a disturbance parameter. It is also possible to show that conditioned to any ϕ , the value for $\boldsymbol{\beta}$ that maximizes $L(\cdot, \cdot)$ does not depend on ϕ . Therefore, $\boldsymbol{\beta}$ and ϕ can be estimated separately.

Let's indicate with $\tilde{\boldsymbol{\beta}}$ the maximum likelihood estimator for $\boldsymbol{\beta}$. Its determination $\hat{\boldsymbol{\beta}}$ is defined as:

$$\hat{\boldsymbol{\beta}} = \arg \max_{\boldsymbol{\beta} \in \mathbb{R}^{p+1}} L(\boldsymbol{\beta}, \phi; \mathbf{y}) \quad (2.3)$$

Finding the values $\hat{\boldsymbol{\beta}}$ that maximize the likelihood corresponds to finding the values that maximize the log-likelihood $l(\boldsymbol{\beta}, \phi; \mathbf{y}) = \log(L(\boldsymbol{\beta}, \phi; \mathbf{y}))$. For the independence hypothesis on Y_1, \dots, Y_n we get:

$$\begin{aligned} l(\boldsymbol{\beta}, \phi; \mathbf{y}) &= \log(L(\boldsymbol{\beta}, \phi; \mathbf{y})) \\ &= \log \left(\prod_{i=1}^n \exp \left\{ \frac{\omega_i}{\phi} [y_i \theta_i - b(\theta_i)] \right\} c(y_i, \phi, \omega_i) \right) \\ &= \sum_{i=1}^n \left\{ \frac{\omega_i}{\phi} [y_i \theta_i - b(\theta_i)] + \log(c(y_i, \phi, \omega_i)) \right\} \\ &= \sum_{i=1}^n l_i(\boldsymbol{\beta}, \phi; \mathbf{y}) \end{aligned} \quad (2.4)$$

The maximum value of $l(\boldsymbol{\beta}, \phi; \mathbf{y})$ can be obtained by imposing all its partial derivatives equal to 0:

$$\frac{\partial l(\boldsymbol{\beta}, \phi; \mathbf{y})}{\partial \beta_j} = 0, \quad \forall j \in \{0, 1, \dots, p\}$$

These equations can be solved with numerical methods, such as Newton-Raphson algorithm or its variant Fisher scoring. It is possible to show that the Fisher scoring algorithm corresponds to iteratively solving a weighted least squares optimization problem.

A statistic that can be used to measures the goodness of fit of a model is the *Deviance*. It can be used by comparing the current model log-likelihood $l(\hat{\boldsymbol{\beta}}, \phi; \mathbf{y})$ with the *saturated model* log-likelihood $l_S(\boldsymbol{\beta}^*, \phi; \mathbf{y})$. The saturated model is the model with n parameter, so a model where the expected values of the response variables μ_1, \dots, μ_n are estimated with their observed values y_1, \dots, y_n . It is possible to show that $l_S(\boldsymbol{\beta}^*, \phi; \mathbf{y}) \geq l(\hat{\boldsymbol{\beta}}, \phi; \mathbf{y})$. The closer $l(\hat{\boldsymbol{\beta}}, \phi; \mathbf{y})$ is to $l_S(\boldsymbol{\beta}^*, \phi; \mathbf{y})$, the better the current model fitting is.

Definition 2.2 (Deviance). Given $l(\hat{\boldsymbol{\beta}}, \phi; \mathbf{y})$ the log-likelihood of the current model and $l_S(\boldsymbol{\beta}^*, \phi; \mathbf{y})$ the log-likelihood of the saturated model, the *Scaled Deviance* of the current model is defined as:

$$S(\hat{\boldsymbol{\beta}}, \phi, \mathbf{y}) = -2 \left(l(\hat{\boldsymbol{\beta}}, \phi; \mathbf{y}) - l_S(\boldsymbol{\beta}^*, \phi; \mathbf{y}) \right)$$

The *Deviance* of the current model is defined as:

$$D(\hat{\boldsymbol{\beta}}, \mathbf{y}) = \phi S(\hat{\boldsymbol{\beta}}, \phi, \mathbf{y})$$

In deviance notation $D(\hat{\boldsymbol{\beta}}, \mathbf{y})$, the parameter ϕ is not reported because the deviance does not depend on ϕ . Indeed, from (2.4) we get:

$$\begin{aligned}
 S(\hat{\boldsymbol{\beta}}, \phi, \mathbf{y}) &= -2 \left(l(\hat{\boldsymbol{\beta}}, \phi; \mathbf{y}) - l_S(\boldsymbol{\beta}^*, \phi; \mathbf{y}) \right) \\
 &= -2 \left(\sum_{i=1}^n \left\{ \frac{\omega_i}{\phi} [y_i \hat{\theta}_i - b(\hat{\theta}_i)] + \log(c(y_i, \phi, \omega_i)) \right\} \right. \\
 &\quad \left. - \sum_{i=1}^n \left\{ \frac{\omega_i}{\phi} [y_i \theta_i^* - b(\theta_i^*)] + \log(c(y_i, \phi, \omega_i)) \right\} \right) \\
 &= -2 \left(\sum_{i=1}^n \frac{\omega_i}{\phi} \{ [y_i \hat{\theta}_i - b(\hat{\theta}_i)] - [y_i \theta_i^* - b(\theta_i^*)] \} \right) \\
 D(\hat{\boldsymbol{\beta}}, \mathbf{y}) &= -2 \left(\sum_{i=1}^n \omega_i \{ [y_i \hat{\theta}_i - b(\hat{\theta}_i)] - [y_i \theta_i^* - b(\theta_i^*)] \} \right)
 \end{aligned}$$

In table 2.3 the deviances for the families mentioned in 2.1 are reported.

Table 2.3: Deviance for Linear Exponential Families

Distribution	Deviance $D(\hat{\boldsymbol{\beta}}, \mathbf{y})$
Normal	$\sum_{i=1}^n (y_i - \hat{\mu}_i)^2$
Poisson	$2 \sum_{i=1}^n \left\{ y_i \log \left(\frac{y_i}{\hat{\mu}_i} \right) - (y_i - \hat{\mu}_i) \right\}$
Gamma	$2 \sum_{i=1}^n \left\{ -\log \left(\frac{y_i}{\hat{\mu}_i} \right) + \frac{y_i - \hat{\mu}_i}{\hat{\mu}_i} \right\}$
Scaled Binomial	$2 \sum_{i=1}^n \left\{ y_i \log \left(\frac{y_i}{\hat{\mu}_i} \right) + (1 - y_i) \log \left(\frac{1 - y_i}{1 - \hat{\mu}_i} \right) \right\}$

As $l_S(\boldsymbol{\beta}^*, \phi; \mathbf{y})$ does not depends on $\hat{\boldsymbol{\beta}}$, maximizing the likelihood in equation (2.3) is the same as minimizing the deviance, that can be seen as a *Loss Function*.

2.1.1.4 Variable effects

2.1.1.5 Variable selection

2.1.2 Elastic Net

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Vivamus id mauris interdum, malesuada ante eu, tempus lacus. Aliquam blandit tortor a velit ultricies, eget pharetra nulla egestas. Suspendisse pellentesque finibus est, vitae ullamcorper magna convallis ut. Nulla a lectus in ligula iaculis convallis. Pellentesque tortor mauris, tempor nec dictum et, facilisis sit amet dolor. Mauris nibh quam, molestie non ex quis, hendrerit dignissim nulla. Aliquam sit amet dui at diam vestibulum malesuada a id lacus. Phasellus viverra orci vitae sem pretium, eu consequat libero euismod.

Cras suscipit aliquam consequat. Quisque sodales lacus ac erat malesuada, eu laoreet enim vestibulum. Sed id ante id ligula auctor ullamcorper. Sed luctus rutrum mollis. Vestibulum sed ultrices quam. Duis id orci ut enim elementum maximus id quis justo. Pellentesque rutrum ligula in aliquam rhoncus. Integer suscipit nisl at mi efficitur interdum. Aenean et orci elit.

Nam ultricies est et iaculis tempus. Quisque leo lorem, sagittis et ligula a, blandit mattis velit. Phasellus pretium, orci et semper finibus, dui nulla tempor nisl, vel vehicula magna diam nec sem. Praesent finibus commodo enim non laoreet. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Curabitur ut pellentesque purus. Proin hendrerit, odio vel sodales porta, ex lorem feugiat sem, non fringilla libero ex ac ligula. Quisque facilisis eros at suscipit rhoncus.

2.1.3 GAM

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Vivamus id mauris interdum, malesuada ante eu, tempus lacus. Aliquam blandit tortor a velit ultricies, eget pharetra nulla egestas. Suspendisse pellentesque finibus est, vitae ullamcorper magna convallis ut. Nulla a lectus in ligula iaculis convallis. Pellentesque tortor mauris, tempor nec dictum et, facilisis sit amet dolor. Mauris nibh quam, molestie non ex quis, hendrerit dignissim nulla. Aliquam sit amet dui at diam vestibulum malesuada a id lacus. Phasellus viverra orci vitae sem pretium, eu consequat libero euismod.

Cras suscipit aliquam consequat. Quisque sodales lacus ac erat malesuada, eu laoreet enim vestibulum. Sed id ante id ligula auctor ullamcorper. Sed luctus rutrum mollis. Vestibulum sed ultrices quam. Duis id orci ut enim elementum maximus id quis justo. Pellentesque rutrum ligula in aliquam rhoncus. Integer suscipit nisl at mi efficitur interdum. Aenean et orci elit.

Nam ultricies est et iaculis tempus. Quisque leo lorem, sagittis et ligula a, blandit mattis velit. Phasellus pretium, orci et semper finibus, dui nulla tempor nisl, vel vehicula magna diam nec sem. Praesent finibus commodo enim non laoreet. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Curabitur ut pellentesque purus. Proin hendrerit, odio vel sodales porta, ex lorem feugiat sem, non fringilla libero ex ac ligula. Quisque facilisis eros at suscipit rhoncus.

2.1.4 GBM

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Vivamus id mauris interdum, malesuada ante eu, tempus lacus. Aliquam blandit tortor a velit ultricies, eget pharetra nulla egestas. Suspendisse pellentesque finibus est, vitae ullamcorper magna convallis ut. Nulla a lectus in ligula iaculis convallis. Pellentesque tortor mauris, tempor nec dictum et, facilisis sit amet dolor. Mauris nibh quam, molestie non ex quis, hendrerit dignissim nulla. Aliquam sit amet dui at diam vestibulum malesuada a id lacus. Phasellus viverra orci vitae sem pretium, eu consequat libero euismod.

Cras suscipit aliquam consequat. Quisque sodales lacus ac erat malesuada, eu laoreet enim vestibulum. Sed id ante id ligula auctor ullamcorper. Sed luctus rutrum mollis. Vestibulum sed ultrices quam. Duis id orci ut enim elementum maximus id quis justo. Pellentesque rutrum ligula in aliquam rhoncus. Integer suscipit nisl at mi efficitur interdum. Aenean et orci elit.

Nam ultricies est et iaculis tempus. Quisque leo lorem, sagittis et ligula a, blandit mattis velit. Phasellus pretium, orci et semper finibus, dui nulla tempor nisl, vel vehicula magna diam nec sem. Praesent finibus commodo enim non laoreet. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Curabitur ut pellentesque purus. Proin hendrerit, odio vel sodales porta, ex lorem feugiat sem, non fringilla libero ex ac ligula. Quisque facilisis eros at suscipit rhoncus.

2.2 Model comparison

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Vivamus id mauris interdum, malesuada ante eu, tempus lacus. Aliquam blandit tortor a velit ultricies, eget pharetra nulla egestas. Suspendisse pellentesque finibus est, vitae ullamcorper magna convallis ut. Nulla a lectus in ligula iaculis convallis. Pellentesque tortor mauris, tempor nec dictum et, facilisis sit amet dolor. Mauris nibh quam, molestie non ex quis, hendrerit dignissim nulla. Aliquam sit amet dui at diam vestibulum malesuada a id lacus. Phasellus viverra orci vitae sem pretium, eu consequat libero euismod.

Cras suscipit aliquam consequat. Quisque sodales lacus ac erat malesuada, eu laoreet enim vestibulum. Sed id ante id ligula auctor ullamcorper. Sed luctus rutrum mollis. Vestibulum sed ultrices quam. Duis id orci ut enim elementum maximus id quis justo. Pellentesque rutrum ligula in aliquam rhoncus. Integer suscipit nisl at mi efficitur interdum. Aenean et orci elit.

Nam ultricies est et iaculis tempus. Quisque leo lorem, sagittis et ligula a, blandit mattis velit. Phasellus pretium, orci et semper finibus, dui nulla tempor nisl, vel vehicula magna diam nec sem. Praesent finibus commodo enim non laoreet. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Curabitur ut pellentesque purus. Proin hendrerit, odio vel sodales porta, ex lorem feugiat sem, non fringilla libero ex ac ligula. Quisque facilisis eros at suscipit rhoncus.

2.3 The actuary importance

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Vivamus id mauris interdum, malesuada ante eu, tempus lacus. Aliquam blandit tortor a velit ultricies, eget pharetra nulla egestas. Suspendisse pellentesque finibus est, vitae ullamcorper magna convallis ut. Nulla a lectus in ligula iaculis convallis. Pellentesque tortor mauris, tempor nec dictum et, facilisis sit amet dolor. Mauris nibh quam, molestie non ex quis, hendrerit dignissim nulla. Aliquam sit amet dui at diam vestibulum malesuada a id lacus. Phasellus viverra orci vitae sem pretium, eu consequat libero euismod.

Cras suscipit aliquam consequat. Quisque sodales lacus ac erat malesuada, eu laoreet enim vestibulum. Sed id ante id ligula auctor ullamcorper. Sed luctus rutrum mollis. Vestibulum sed ultrices quam. Duis id orci ut enim elementum maximus id quis justo. Pellentesque rutrum ligula in aliquam rhoncus. Integer suscipit nisl at mi efficitur interdum. Aenean et orci elit.

Nam ultricies est et iaculis tempus. Quisque leo lorem, sagittis et ligula a, blandit mattis velit. Phasellus pretium, orci et semper finibus, dui nulla tempor nisl, vel vehicula magna diam nec sem. Praesent finibus commodo enim non laoreet. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Curabitur ut pellentesque purus. Proin hendrerit, odio vel sodales porta, ex lorem feugiat sem, non fringilla libero ex ac ligula. Quisque facilisis eros at suscipit rhoncus.

2.4 Implementation

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Vivamus id mauris interdum, malesuada ante eu, tempus lacus. Aliquam blandit tortor a velit ultricies, eget pharetra nulla egestas. Suspendisse pellentesque finibus est, vitae ullamcorper magna convallis ut. Nulla a lectus in ligula iaculis convallis. Pellentesque tortor mauris, tempor nec dictum et, facilisis sit amet dolor. Mauris nibh quam, molestie non ex quis, hendrerit dignissim nulla. Aliquam sit amet dui at diam vestibulum malesuada a id lacus. Phasellus viverra orci vitae sem pretium, eu consequat libero euismod.

Cras suscipit aliquam consequat. Quisque sodales lacus ac erat malesuada, eu laoreet enim vestibulum. Sed id ante id ligula auctor ullamcorper. Sed luctus rutrum mollis. Vestibulum sed ultrices quam. Duis id orci ut enim elementum maximus id quis justo. Pellentesque rutrum ligula in aliquam rhoncus. Integer suscipit nisl at mi efficitur interdum. Aenean et orci elit.

Nam ultricies est et iaculis tempus. Quisque leo lorem, sagittis et ligula a, blandit mattis velit. Phasellus pretium, orci et semper finibus, dui nulla tempor nisl, vel vehicula magna diam nec sem. Praesent finibus commodo enim non laoreet. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Curabitur ut pellentesque

purus. Proin hendrerit, odio vel sodales porta, ex lorem feugiat sem, non fringilla libero ex ac ligula. Quisque facilisis eros at suscipit rhoncus.

→

3

Practical application

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Vivamus id mauris interdum, malesuada ante eu, tempus lacus. Aliquam blandit tortor a velit ultricies, eget pharetra nulla egestas. Suspendisse pellentesque finibus est, vitae ullamcorper magna convallis ut. Nulla a lectus in ligula iaculis convallis. Pellentesque tortor mauris, tempor nec dictum et, facilisis sit amet dolor. Mauris nibh quam, molestie non ex quis, hendrerit dignissim nulla. Aliquam sit amet dui at diam vestibulum malesuada a id lacus. Phasellus viverra orci vitae sem pretium, eu consequat libero euismod.

3.1 Data description

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Vivamus id mauris interdum, malesuada ante eu, tempus lacus. Aliquam blandit tortor a velit ultricies, eget pharetra nulla egestas. Suspendisse pellentesque finibus est, vitae ullamcorper magna convallis ut. Nulla a lectus in ligula iaculis convallis. Pellentesque tortor mauris, tempor nec dictum et, facilisis sit amet dolor. Mauris nibh quam, molestie non ex quis, hendrerit dignissim nulla. Aliquam sit amet dui at diam vestibulum malesuada a id lacus. Phasellus viverra orci vitae sem pretium, eu consequat libero euismod.

Cras suscipit aliquam consequat. Quisque sodales lacus ac erat malesuada, eu laoreet enim vestibulum. Sed id ante id ligula auctor ullamcorper. Sed luctus rutrum mollis. Vestibulum sed ultrices quam. Duis id orci ut enim elementum maximus id quis justo. Pellentesque rutrum ligula in aliquam rhoncus. Integer suscipit nisl at mi efficitur interdum. Aenean et orci elit.

Nam ultricies est et iaculis tempus. Quisque leo lorem, sagittis et ligula a, blandit mattis velit. Phasellus pretium, orci et semper finibus, dui nulla tempor nisl, vel vehicula magna diam nec sem. Praesent finibus commodo enim non laoreet. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Curabitur ut pellentesque purus. Proin hendrerit, odio vel sodales porta, ex lorem feugiat sem, non fringilla libero ex ac ligula. Quisque facilisis eros at suscipit rhoncus.

3.2 Model used

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Vivamus id mauris interdum, malesuada ante eu, tempus lacus. Aliquam blandit tortor a velit ultricies, eget pharetra nulla egestas. Suspendisse pellentesque finibus est, vitae ullamcorper magna convallis ut. Nulla a lectus in ligula iaculis convallis. Pellentesque tortor mauris, tempor nec dictum et, facilisis sit amet dolor. Mauris nibh quam, molestie non ex quis, hendrerit dignissim nulla. Aliquam sit amet dui at diam vestibulum malesuada a id lacus. Phasellus viverra orci vitae sem pretium, eu consequat libero euismod.

Cras suscipit aliquam consequat. Quisque sodales lacus ac erat malesuada, eu laoreet enim vestibulum. Sed id ante id ligula auctor ullamcorper. Sed luctus rutrum mollis. Vestibulum sed ultrices quam. Duis id orci ut enim elementum maximus id quis justo. Pellentesque rutrum ligula in aliquam rhoncus. Integer suscipit nisl at mi efficitur interdum. Aenean et orci elit.

Nam ultricies est et iaculis tempus. Quisque leo lorem, sagittis et ligula a, blandit mattis velit. Phasellus pretium, orci et semper finibus, dui nulla tempor nisl, vel vehicula magna diam nec sem. Praesent finibus commodo enim non laoreet. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Curabitur ut pellentesque

purus. Proin hendrerit, odio vel sodales porta, ex lorem feugiat sem, non fringilla libero ex ac ligula. Quisque facilisis eros at suscipit rhoncus.

3.3 Model assessment

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Vivamus id mauris interdum, malesuada ante eu, tempus lacus. Aliquam blandit tortor a velit ultricies, eget pharetra nulla egestas. Suspendisse pellentesque finibus est, vitae ullamcorper magna convallis ut. Nulla a lectus in ligula iaculis convallis. Pellentesque tortor mauris, tempor nec dictum et, facilisis sit amet dolor. Mauris nibh quam, molestie non ex quis, hendrerit dignissim nulla. Aliquam sit amet dui at diam vestibulum malesuada a id lacus. Phasellus viverra orci vitae sem pretium, eu consequat libero euismod.

Cras suscipit aliquam consequat. Quisque sodales lacus ac erat malesuada, eu laoreet enim vestibulum. Sed id ante id ligula auctor ullamcorper. Sed luctus rutrum mollis. Vestibulum sed ultrices quam. Duis id orci ut enim elementum maximus id quis justo. Pellentesque rutrum ligula in aliquam rhoncus. Integer suscipit nisl at mi efficitur interdum. Aenean et orci elit.

Nam ultricies est et iaculis tempus. Quisque leo lorem, sagittis et ligula a, blandit mattis velit. Phasellus pretium, orci et semper finibus, dui nulla tempor nisl, vel vehicula magna diam nec sem. Praesent finibus commodo enim non laoreet. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Curabitur ut pellentesque purus. Proin hendrerit, odio vel sodales porta, ex lorem feugiat sem, non fringilla libero ex ac ligula. Quisque facilisis eros at suscipit rhoncus.

3.4 Results

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Vivamus id mauris interdum, malesuada ante eu, tempus lacus. Aliquam blandit tortor a velit ultricies, eget pharetra nulla egestas. Suspendisse pellentesque finibus est, vitae ullamcorper magna convallis ut. Nulla a lectus in ligula iaculis convallis. Pellentesque tortor mauris, tempor nec dictum et, facilisis sit amet dolor. Mauris nibh quam, molestie

non ex quis, hendrerit dignissim nulla. Aliquam sit amet dui at diam vestibulum malesuada a id lacus. Phasellus viverra orci vitae sem pretium, eu consequat libero euismod.

Cras suscipit aliquam consequat. Quisque sodales lacus ac erat malesuada, eu laoreet enim vestibulum. Sed id ante id ligula auctor ullamcorper. Sed luctus rutrum mollis. Vestibulum sed ultrices quam. Duis id orci ut enim elementum maximus id quis justo. Pellentesque rutrum ligula in aliquam rhoncus. Integer suscipit nisl at mi efficitur interdum. Aenean et orci elit.

Nam ultricies est et iaculis tempus. Quisque leo lorem, sagittis et ligula a, blandit mattis velit. Phasellus pretium, orci et semper finibus, dui nulla tempor nisl, vel vehicula magna diam nec sem. Praesent finibus commodo enim non laoreet. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Curabitur ut pellentesque purus. Proin hendrerit, odio vel sodales porta, ex lorem feugiat sem, non fringilla libero ex ac ligula. Quisque facilisis eros at suscipit rhoncus.

This is my bibliography

Bibliography

- Darwin, Charles (1859). “On the Origin of Species by Means of Natural Selection or the Preservation of Favoured Races in the Struggle for Life”. London: John Murray.
- Lottridge, Danielle, Eli Marschner, Ellen Wang, Maria Romanovsky, and Clifford Nass (2012). “Browser design impacts multitasking”. In: *Proceedings of the Human Factors and Ergonomics Society 56th Annual Meeting*.
- Satchell, Christine and Paul Dourish (2009). “[Beyond The User: Use And Non-Use in HCI](#)”. In: *Proceedings of the Annual Conference of the Australian Computer-Human Interaction Special Interest Group (OZCHI '09)* November, pp. 9–16.
- Shea, Nicholas, Annika Boldt, Dan Bang, Nick Yeung, Cecilia Heyes, and Chris D Frith (2014). “[Supra-personal cognitive control and metacognition](#)”. In: *Trends in Cognitive Sciences* 18.4, pp. 186–193.
- Von Goethe, Johann Wolfgang (1829). “Wilhelm Meisters Wanderjahre oder die Entsagenden”. de. Cotta.
- Wu, Tim (2016). “The Attention Merchants: The Epic Scramble to Get Inside Our Heads”. Knopf Publishing Group.