Application of GLM Advancements  
to Non-Life Insurance Pricing

Leonardo Stincone

This is my abstract

Table of Contents

# 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 we are 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:

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 caused 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:

* , with , be the coverage period;
* be the premium payed by the policyholder to the insurer;
* be the number of claims occurred during the coverage period (*claims count*);
* , with , be the timing of each claim;
* be the amount of each claim (*claims severities* or *claims sizes*).

The total cost of claims for the insurance is:

For simplicity, in the following we are going to just use the notation with the meaning of if .

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 with one single negative cash flow . On the other hand, the insurer undertakes the negative cash flows in exchange for a positive cash flow .

The major difference between these cash flows is that is a certain amount, while , at the time , are uncertain in the amount, in the count () and in the timing (). So, the policyholder, paying a premium , 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 , while the costs will emerge later. In most 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 predict the future costs in order to determine an adequate premium.

From a statistical point of view, we can translate this uncertainty saying that and are random variables. Therefore, we can say that is a stochastic process. Usually, in non-file insurance pricing, the variables are not taken into account because the coverage span is short and from a financial point of view the timing of the claims occurrences has negligible effect.

Previously we said that are all positive. 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 with we can use the rule that , so . Therefore, we can say that:

![Figure 1.1: Insurance Contract cash flows.](data:application/pdf;base64,)

Figure 1.1: Insurance Contract cash flows.

## 1.2 Non-Life insurance pricing

In insurances, the premium that the 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 for a specific policyholder , given the known information on him. In the next sections we are going to better explain what “proper” means.

### 1.2.1 Compound distribution hypotheses

The first step for evaluating the stochastic process 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 , the variables are stochastically independent and identically distributed;
2. the probability distribution of does not depend on .

Under these hypotheses we say that:

has a compound distribution.

The variable used in this definition can be interpreted as the *claim severity for the th claim under the hypothesis that claims occurred*. The two hypotheses provided in definition 1.2 imply that the distribution of does not depend on nor on . For this reason, in the following, we are going to use the notation to represent a random variable with the ditribution of and for its cumulative distribution function (i.e. ).

Let’s consider the variabile . We can interpret it as the *claim severity for the th claim under the hypothesis that the th claim occurred*. From the hypotheses provided in definition 1.2 we can obtain that also has the same distribution of . This can be easily obtained as follows:

Where:

* the step (??) and the step (??) are given by the fact that the event can be decomposed as and that the events are two-by-two disjoint, so they constitute a partition of , that allows us to use the disintegrability property of the probability;
* the step (??) is due to the fact that the distribution of depends neither on nor on ;
* the equivalence at step (??) is due to the fact that the events are a partition of .

Therefore, 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 given the distribution of and . In this chapter we are going to provide only the formula of the expected value , but, with the same approach one can obtain all the moments.

The expected value of the total cost of claims can be obtained from the expected value of the claims count and the expected value of the claim severity as follows:

Where:

* the step (??) is given by the fact that the events constitute a partition of the certain event , that allows us to use the disintegrability property of the expected value;
* the step (??) is due to the definition of ;
* the step (??) is due to the linearity of the expected value;
* the steps (??) and (??) are due to the fact that, as assumed by the compound distribution hypotheses, does not depends on and ;
* the step (??) is due to the definition of the expected value .

This result tells us that, under the hypotheses of the compound distribution, it is possible to easily obtain from and . That means that we can model separately and and, from them, obtain . That result is particularly useful in personalization (paragraph 1.3), because, for each individual , given the information we have on him, we can estimate his expected claim size and his expected claim severity and obtain his expected total cost of claims as .

### 1.2.3 Risk premium and Technical Price

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

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

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 the total cost of claims of a policyholder and the expenses related to his policy, his *Technical Price* is given by:

Secondly, even if the policyholder paid 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 amount of the Technical Price with these loadings can be further modified based on business logic, as we are 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](https://www.ania.it/)) 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. . 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 has only 2 possible modalities, it can be easily encoded in a binary variable that assigns to one modality and to the other. For example, if , it can be encoded this way:

In general, if a variable has modalities, it can be encoded in binary variables . For example, if and it can have 4 possible modalities (‘Fiat’, ‘Alfa-Romeo’, ‘Lancia’, ‘Ferrari’) it can be encoded this way:

The variables , , 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:

In table 1.1 the dummy variable encoding is illustrated.

Table 1.1: Dummy variables encoding.

Make

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 th modality. This encoding is called one-hot encoding and it is mainly used in Neural Networks. For the models considered in this paper the dummy variables encoding is preferred, so we will always consider it.

In the following, when we use the notation , I’ll always consider that the qualitative variables have been already encoded as dummy variables, so

### 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 that from an instance of a set of pricing variables returns a price:

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 and obtain a pricing rule for the risk premium .

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

that, given a set of explanatory variables , returns the expected value of the response variable and possibly other moments of or even the full distribution of .

In definition 1.6 we used a generic as codomain of the function to not specify whether the model describes just (and so ) or something more, such as the couple or the full distribution of .

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 , but we can separately model and .

### 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

The claim count is a discrete variable with values in . Even if in practice the number of claims can’t be arbitrarily high, is usually modeled with distributions that give a positive probability to all natural numbers. One of the most common distribution used for is the Poisson distribution.

Definition 1.7: (Poisson Distribution) A random variable with support has a Poisson distribution, if its probability function is:

We will indicate it with the notation .

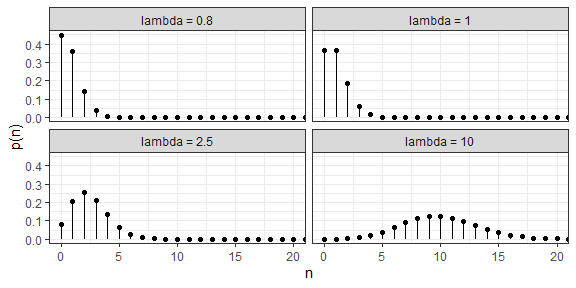


Figure 1.2: Poisson distribution for some values of .

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

Indeed, the first two moments are:

Thus, increasing , both and increase.

Looking to the distribution shape, we can see that:

* if , the mode is in ;
* if , ;
* if , the mode is in a value greater than and, as 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 . E.g. the average number of claims for motor third party liability insurances in Italy, in 2018 has been 5.68%[[1]](#footnote-46).

The property 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 . 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 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 spans in two consecutive years and , it is collected as two records: the couple and the couple . 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 . 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 the number of claims the policyholder will experience during his period of exposure (measured in years) and the number of claims the policyholder would experience during one year, we assume .

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

Definition 1.8: (Counting Process) A stocastic process is called if:

In a counting process :

* can be interpreted as the number of events or arrivals that occur in the period ;
* can be interpreted as the number of events or arrivals that occur in the period . 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 is a with intensity if:

Under these hypotheses we obtain the following result:

Theorem 1.1: (Poisson Process) If is a Poisson process with intensity , then:

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

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

If we assume that the claims that occur to a policy come from a Poisson process with intensity , if we observe that policy for the period , the claims count in that exposure period are distributed as:

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

#### 1.3.4.3 Distribution for the claim severity

The claim severity is a continuous variable with values in . As for the claims count , even if in practice it can’t be arbitrarily high, it is usually modeled with distributions that give a positive density to all the numbers in . As the null claims are excluded, it is natural to assume . One of the most common distribution used for is the Gamma distribution.

Definition 1.10: (Gamma Distribution) A random variable with support has a Gamma distribution, if its probability density function is:

where .

We will indicate it with the notation .

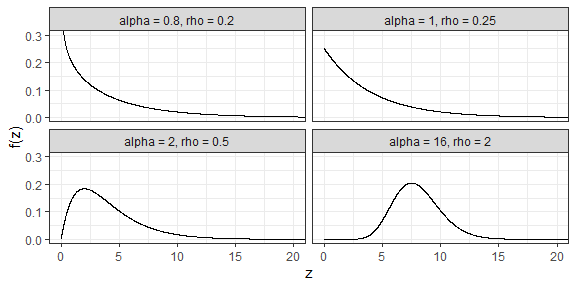


Figure 1.3: Gamma distribution for some values of and .

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

* , called shape parameter
* , called scale parameter

The first two moments of the Gamma distribution are:

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

* if , is not defined in and it has a vertical asymptote in . In it is monotonically decreasing.
* if , starts from and then decreases monotonically. In this case, the density function becomes and the distribution is also called exponential distribution.
* if , starts from , increases until the mode and then decreases.

In figure 1.3 the first three distributions represented have the same expected value , but different shapes. The third and the fourth have the same variance , but different expected values. As the shape parameter 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 as a parameter:

With this parametrization, the density function becomes:

The advantage of using the parameters is that the link between and becomes clearer:

Computing the coefficient of variation we then obtain:

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

Another characteristic of Gamma distribution that could be problematic in modeling claims severity is that it has a light tail. This means that, as goes to , aproaches 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 and considering large all the claims with a size that exceeds that threshold. The value must be chosen sufficiently big to consider large the claims above , but not so big that there are not enough observed claims that exceeds . One common choice for Motor Third Party Liability in European markets could be $\bar{z} = 100' 000 \text{\euro}$.

Definition 1.11: (Large and Attritional Claims) Given a predetermined threshold , we say that:

For each claim we call:

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

![Figure 1.4: Large claims.](data:application/pdf;base64,)

Figure 1.4: Large claims.

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

where

* is the attritional claims count, i.e. the number of claims with size ;
* is the large claims count, i.e. the number of claims with size ;

Let’s indicate with the th in order from the smallest to the bigger. Sorting the claims we can separate the attritional claims from the large claims as follows:

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

These three decompositions of are useful because they provide three decompositions of :

??, ?? and ?? provide three approaches to model attritional and large claims.

1. Looking to ?? we can model separately attritional claims and large claims. Modeling and we estimate the total cost of claims for the attritional part ; modeling and we estimate the total cost of claims for the large part .
2. Looking to ?? we can model together the claim count , and then we can model the cost of the attritional claims , the cost of the large claims and the probability to exceed the threshold .
3. Looking to ?? we can model together the claim count , and then we can model the capped claims size , the excess over the threshold and the probability to exceed the threshold .

If the large claims component weighs a lot on the total cost of claims, these approaches could lead to quite different estimates of . In particular, if in the observed data the number of large claims is small, it will be hard to model both and , so for these components the modeling process could lead to a flat model (i.e. a model without any explanatory variable) or almost flat one (i.e. a model with just few explanatory variables and with mild effects). However, with the first approach, a flat model for 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 leads to distribute the observed total cost of large claims proportionally to the expected number of claims . 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 .

For the second approach we must also introduce a distribution suitable for modeling .

#### 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 . For example we can use it to model the number of large claims within a fixed number of claims.

Definition 1.12: (Binomial Distribution) A random variable with support has a Binomial distribution, if its probability function is:

We will indicate it with the notation .

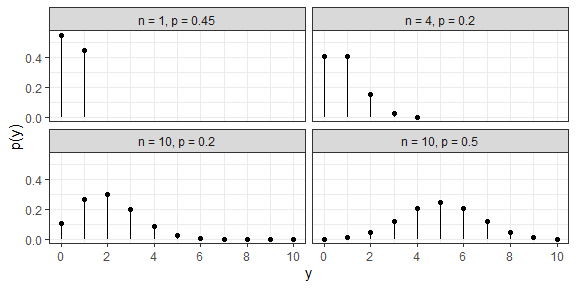


Figure 1.5: Binomial distribution for some values of and .

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

The first two moments of the binomial distribution are:

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

If , the binomial distribution assumes a shape centered on its expected value and fading for values of that moves away from . As 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 .

Definition 1.13: (Scaled Binomial Distribution) If , and , we will say that has a and we will indicate it with the notation .

The support of is and its probability function is:

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, , in order to model separately attritional claims severity and large claims severity , as we have seen in 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 in the following two ways:

where:

* is the number of claims with only material damages;
* is the number of claims with injuries;
* is the event that represents that a specific claim presents injuries; such as is a representative for , is a representative for .

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:

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 , we have to model the response variables needed for that decomposition (, , , …) with the explanatory variables. Thus we have to estimate a function as defined in 1.6.

In order to estimate we have also to take some assumptions on the distribution of the response variable and on the shape of . We will call *model* a set of assumptions on the response variable and on the shape of . 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 with the explanatory variables , the observed data is in the form:

where:

* is the number of observations in the dataset;
* is the set of explanatory variable for the observation ;
* is the weight for the observation ;
* is the realization of the response variable for the observation .

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

* If we are modeling the yearly claim count , each observation could be a policy (or a couple (policy, accounting year)), the weights could be the exposures 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 , each observation could be a claim , the weights could all be and the realizations of response variables could be the observed cost for the claim . It is also possible to model the claim severity taking into account the total cost of claims for the policy . In this case, each observation would be a policy , the weights would be the number of claims for each policy and the realizations of response variables would be the total observed cost for the claims of the policy .
* If we are modeling the occurrence of injuries in a claim , each observation could be a claim , the weights could be all and the realizations of response variables could be an indicator that assume the value if the claim caused injuries and otherwise. As for the claim severity, we can also aggregate data for policy, so each observation would be a policy , the weights would be the number of claims for the policy and the realizations of response variables would be the number of claims that caused injuries among the claims of the policy .

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

#### 1.3.5.1 Settlement process and IBNR claims

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 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 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 payed. In we have the settlement and in the claim is closed. In is possible that , 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 . These intermediate payments are payed at times . It is also possible that a claim is opened and then gets closed without any payment. After the closing () it is also possible that a claim is reopened and that more payments emerge.

![Figure 1.6: Claim timeline.](data:application/pdf;base64,)

Figure 1.6: Claim timeline.

From the moment the claim is reported (), 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 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 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 . In general 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 and at a specific time , we can’t observe the total number of claims occurred to each policy and the payments for each claim . What we can use is:

* where:
  + is the number of reported claims in for the policy ;
* where:
  + is the amount already payed in for the claim ;
  + is the amount reserved in for the claim .

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

The common case is that and . If we used and without any correction, we would underestimate both and , obtaining a biased estimate for .

To tackle these problems what is usually done is fitting the models for with and and then apply a flat corrective coefficient to based on an aggregated estimate of that takes into account the long settlement process.

An estimate for the expected total cost of claims for a generic policy in the portfolio 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 . Let’s call it . 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 and , the fitting for all the models needed in the decomposition of is performed and, for each policy , is obtained. Let’s call it . As we used the data available in that comes from claims not totally settled, is a biased estimate for .

We can then balance the estimates with by computing:

and applying to the estimates as follows:

We will call rebalanced estimates.

The property of these rebalanced estimates is that on average they are equal to :

So, if is a unbiased estimator for , we obtain:

This procedure can be further developed by balancing not directly the total cost of claims , 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 .

If the dataset contains policies from many years and during the last years a relevant change in the portfolio risk mixture happened, it is also possible to compute only with the policies from the last year of the dataset, rather than with all the policies of the dataset.

The fact that the final estimates are rebalanced on means that the explanatory variables effects estimated with and 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 that is two times the old people one, that relative coefficient 2 will be kept also in the balanced estimate .

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

1. *Tariff Requirement* (or *Fabbisogno Tariffario*): the estimate of by aggregated data;
2. *Personalization*: the estimate of 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*
* the *Technical Price*

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 costraints 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 . The premium obtained this way is called *Offer Price*.

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 customers needs with important discounts on both new business and renewals. From a technical point of view, these discounts have been funded by the remarkable decrese 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 logic. 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 be perceived as more valuable, it can sell it at a higher price than other companies. For example this can be achived by improving customer service and customer experience.

If the technical price and the conversion probability functions 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 conducts the analysis and defines 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]](#footnote-70)

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 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 prices 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 increase 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 implies 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 has some relevant implications on how technical pricing is conducted that we will further discuss in section 2.2.2.

–>

# 2 **Statistical models for Non Life Insurance Pricing**

In this chapter we are going to illustrate some of the most widespread models for technical pricing. For each model we are going to describe its benefits and drawbacks and in section 2.2 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 Generalized Additive Model (GAM) and the Shrinkage estimators for GLM. 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 Families

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 a linear exponential family is and which distributions fit its definition.

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

where:

An exponential family is characterized by the elements . By properly choosing the sets and the functions , 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.

Table 2.1: Some Linear Exponential Families.

Distribution

Notation

Normal

,

Poisson

,

Gamma

,

ScaledBinomial

,

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 . If is a random variable with distribution belonging to an exponential family with parameters , its first two moments are:

As, within a specified family, the parameters and determine a distribution, in practical problems the object of estimation will be the couple . In many problems it is natural to consider distributions from a linear exponential family where the dispersion parameter can be expressed as , where is a known *weight* and is a parameter that we will keep calling *dispersion parameter*. In this case, the density of probability function depends on the parameters and and will be expressed as:

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

* the exposure in the Poisson distribution;
* the number of trials in the Binomial distribution.

#### 2.1.1.2 Model assumptions

Let’s assume that, for statistical units, the observations are available, where is a vector of explanatory variables determinations, is a known weight and is the response variable determination. are all real numbers. The vector is considered a determination of the response random vector .

In GLM we assume that:

1. The response variables are stochastically independent and with probability distribution belonging to a same linear exponential family; i.e. the probability distribution of has density function (or probability function in the discrete case) that can be expressed as:
2. We highlight that only and depend on , while the dispersion parameter is the same for all the observations.
3. The explanatory variables determinations vector affects the probability distribution of the response variable by the linear predictor:
4. that is a linear function of the regression parameters .
5. The linear predictor is linked to the expected value of the response variable by the following relation:
6. where is a monotonic function with continuous first and second derivatives. 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 the design matrix, i.e. the matrix in which each row represents the vector of the explanatory variables for the observation and each column represents the vector of the observations for the explanatory variable . 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 . We assume, as it is common in actuarial datasets, that .

![Figure 2.1: Design Matrix \boldsymbol{X}.](data:application/pdf;base64,)

Figure 2.1: Design Matrix .

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

where must be intended as the vectorial function that links every to .

We assume the design matrix to be a full rank matrix, i.e. . This assumption corresponds to assuming that the columns of are linearly independent.

The function can be chosen as any monotonic function with continuous first and second derivatives. Given a family , a common choice is its canonical link function that is defined as:

From (??) we obtain that, as , choosing the canonical function corresponds to using as the linear predictor:

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

Derivative

Canonical link function

Normal

Poisson

Gamma

ScaledBinomial

In the Gamma case, its canonical function has the drawback that it links the expected values to . This would require some constraints on because would have to be . For this reason, it is preferred to use that maps to .

In the Scaled Binomial case the canonical function is called logit and its inverse is called logistic. For Scaled Binomial distribution we keep using the notation for the expected value as it corresponds to the probability of success .

#### 2.1.1.3 Model fitting

The model depends on the parameters . Indeed, the parameters can be obtained by as:

Therefore, fitting the model corresponds to estimating . The technique used in GLM is the *maximum likelihood*. Let’s indicate with the model likelihood. We remind that the likelihood is a function of the parameters that maps to the density (or probability in the discrete case) of the observed values conditioned to the parameters

The maximum likelihood estimates are the values that maximize . In practice, are the parameters of interest, while is considered as a disturbance parameter. It is possible to show that conditioned to any , the value for that maximizes does not depend on . Therefore, and can be estimated separately.

Let’s indicate with the maximum likelihood estimator for . Its determination is defined as:

$$\begin{equation}
\label{eq:max-lik-est}
\hat{\boldsymbol{\beta}} = \argmax\_{\boldsymbol{\beta}\in\mathbb{R}^{p+1}}{L\left(\boldsymbol{\beta}, \phi; \boldsymbol{y}\right)}
\end{equation}$$

Finding the values that maximize the likelihood corresponds to finding the values that maximize the log-likelihood . For the independence hypothesis on we get:

The maximum value of can be obtained by imposing all its partial derivatives equal to :

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

In the case with Normal response and identity link, the optimization problem (??) has an explicit solution:

A statistic that can be used to measure the goodness of fit of a model is the *Deviance*. It can be used by comparing the current model log-likelihood with the *saturated model* log-likelihood . The saturated model is the model with parameters, so a model where the expected values of the response variables are estimated with their observed values . It is possible to show that . The closer is to , the better the current model fitting is.

Definition 2.2: (Deviance) Given the log-likelihood of the current model and the log-likelihood of the saturated model, the of the current model is defined as:

The of the current model is defined as:

In deviance notation , the parameter is not reported because the deviance does not depend on . Indeed, from (??) we get:

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

Normal

Poisson

Gamma

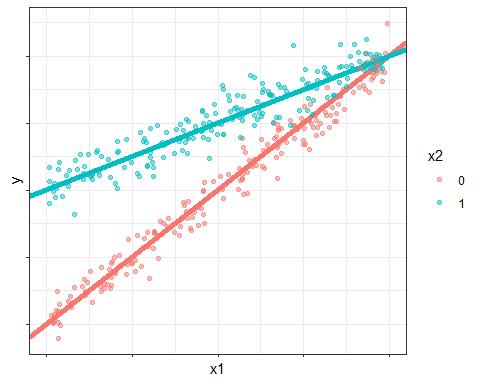
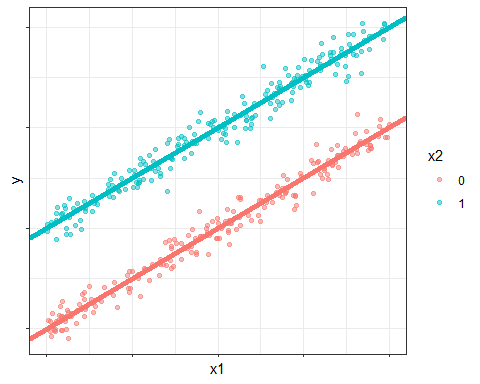
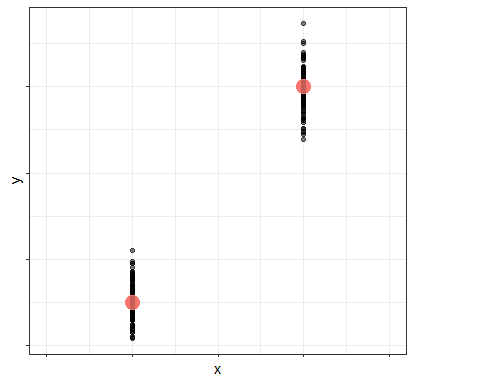
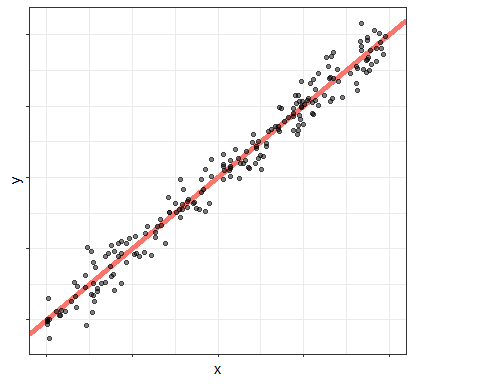
ScaledBinomial

As does not depend on , maximizing the likelihood in equation (??) is the same as minimizing the deviance, that can be seen as a *Loss Function*:

$$\begin{equation}
\label{eq:max-lik-est-deviance}
\hat{\boldsymbol{\beta}} = \argmin\_{\boldsymbol{\beta}\in\mathbb{R}^{p+1}}{D(\boldsymbol{\beta}, \boldsymbol{y})}
\end{equation}$$

#### 2.1.1.4 Variable effects

As we mentioned in 1.3.2, the explanatory variables can be *quantitative* or *qualitative*. In GLM, if explanatory variables transformation terms aren’t added to the linear predictor , the variables effect on is linear. In figure 2.2 the effects of quantitative and qualitative variables are shown. The data is simulated from a GLM with Normal response and identity link.



In the top-left panel, we see the effect of the quantitative variable in the model . As we can see it is a straight line. The coefficient represents the slope of the line, thus means that and are positively correlated, while means that and are negatively correlated. For example, if is the power of the vehicle and the yearly number of claims, means that the more powerful the vehicle is, the more claims the policyholder will experience on average.

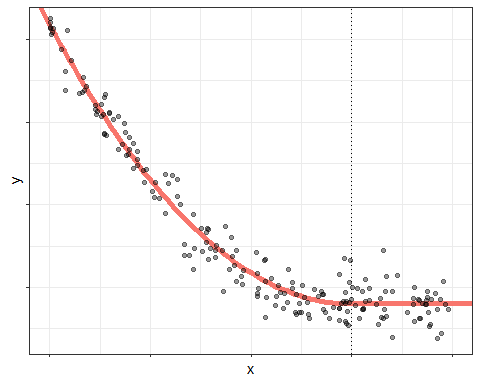
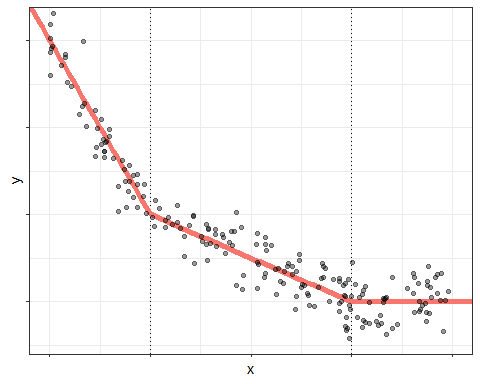
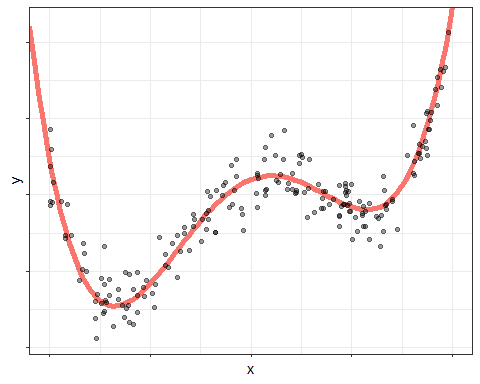
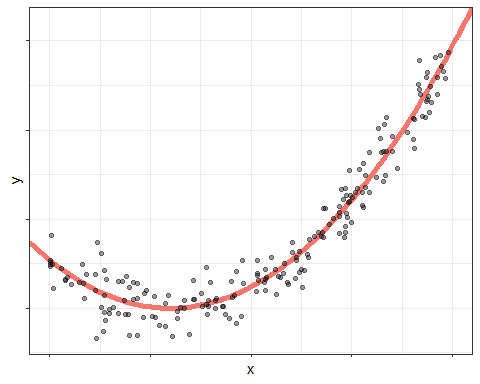
In the top-right panel, we see the effect of a qualitative binary variable in the model . The variable is encoded with values and , so represents the effect of the modality . In general, for a qualitative variable with modalities we will have dummy variables and the model will be . Thus, the coefficient represents the relative effect of the modality compared to the base level modality, that is the one not explicitly included in the dummy encoding. For example, if is the vehicle make, the yearly number of claims, the base level for is ‘Fiat’ and the th modality is ‘Ferrari’, then means that Ferrari cars on average experience more claims that Fiat cars.

In general, in a multivariate model, the coefficient represents the effect of the variable given all the others. For example, in the example of Fiat and Ferrari cars, if in the model there is also the variable ‘vehicle power’, the coefficient corresponding to the modality ‘Ferrari’ represents the how more risky a Ferrari car is compared to a Fiat car with the same power. If the explanatory variables are strongly correlated, it is important to be aware of this aspect. For example, Ferrari cars are usually more powerful that Fiat cars. So, it is possible that in general Ferrari cars are more risky than Fiat cars, but comparing a Ferrari car to a Fiat with the same power, the Ferrari could be less risky. This effect is called *Simpson’s paradox*.[[3]](#footnote-86)

In the bottom-left panel of figure 2.2, we see the effect of a quantitative variable and a qualitative binary variable together in the model . As we can seen, the effects of variable in the two groups defined by variable are represented by two parallel straight lines. The first one is and the second is . The coefficient represents the vertical distance between the two lines.

In the bottom-right panel, the interaction effect between and is included in the model. The model becomes . That means that the effect of variables depends on the determination of the variable. In the group with the effect is represented by the line ; the group with the effect is represented by the line .

For quantitative variables, it is possible to consider also non linear effects in GLMs. Some examples are reported in figure 2.6.



The basic way to achieve it is by adding polynomial terms to the linear predictor. For instance, if is a quantitative variable, it is possible to add to the model the term , obtaining the model . An example of model with both and terms is represented in top-left panel of figure 2.6. Adding the quadratic term, the effect graph becomes a parabola.

With the same logic, it is possible to add more power terms. In general, if we want to model with a polynomial of degree , we can consider the model . In top-right panel of figure 2.6 a 4th degree polynomial effect is represented. We highlight that the model is still considered linear, as the attribute “Linear” in “General Linear Model” is referred to the relation between the parameters and the linear predictor that is still linear.

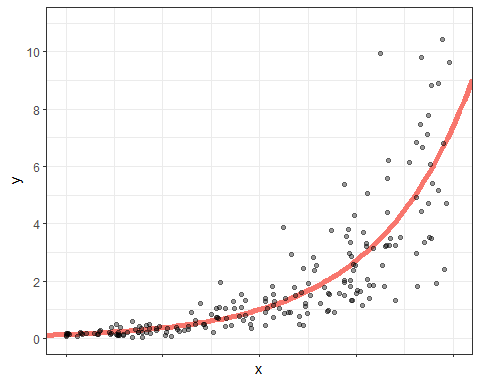
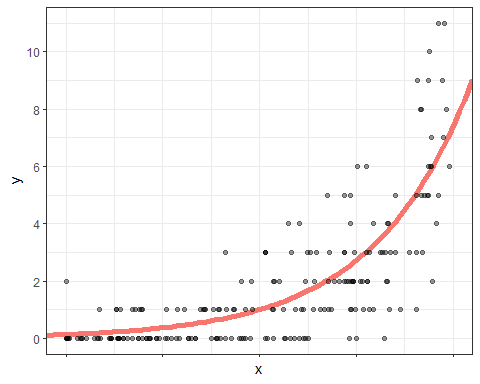
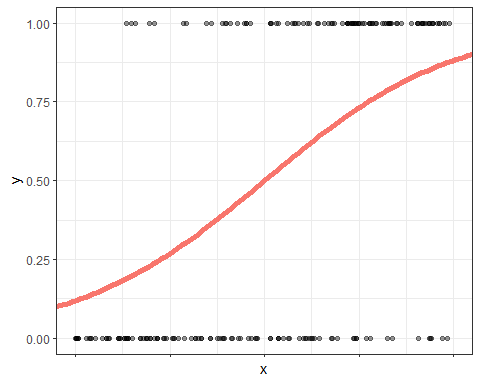
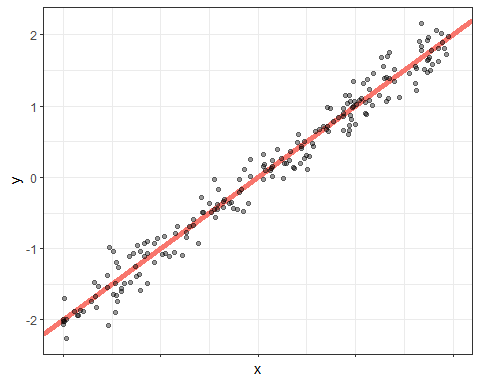
Another way to model non liner effects of explanatory variables is to separate the effects by pieces. In bottom-left panel of figure 2.6 a case in which the effect is separated in 3 pieces is represented. As in all the pieces the effect is linear, the graph of the variable effect is a broken line. This effect can be achieved by adding to the model the terms , where represents the positive part of () and is the value of in the angular point. The values are called *knots*, If the knots are , the model can be represented as . This kind of functions are called *linear splines* and will be further discussed in section 2.1.2. If we want the effect to be null from a certain point , we can consider the variable instead of . This corresponds to aggregate to all the after .

The piece-wise approach can be enhanced by also considering polynomial terms. For instance, in bottom-left panel of figure 2.6, the model represented is , where is the negative part of (). is a parabola with vertex in . The fact of not adding the linear term leads to a monotonic effect made by a semi-parabola and a horizontal semi-line that starts from its vertex.

The examples represented in figures 2.2 and 2.6 are based on simulated data. That means that the linear predictor structure is known and the coefficients are known. In practice, the real model is not known and the coefficients and the structure must be estimated by the data. Thus, we can take assumptions on the structure and we can estimate the coefficients with . In many cases it is not so clear whether to consider or not a variable and how to consider it. For example, with the same data both bottom-left and bottom-right models could work fine. In section 2.1.1.6 we are going to discuss some variable selection techniques for GLM.

#### 2.1.1.5 Link functions and relativities

As we mentioned in 2.1.1.1, GLM supports several families. In figure 2.10 the models with different response variable distributions and link functions are represented. As we can see from the plots, a linear effect on corresponds to a logistic effect when the link is logit and to an exponential effect when the link is log.



If , the model structure can be expressed as:

The term can be seen as the multiplicative factor corresponding to the variable . If is a dummy variable, is the factor the expected value is multiplied by when . If is a quantitative variable, is the factor the expected value is multiplied by for every one-unit increasing of . Indeed:

The fact that with a log link the relation between coefficients and expected value becomes multiplicative is particularly useful to deal with exposure . In section 1.3.4.2 we have seen that often the observations are couples (policy, accounting year), so they have different exposures . Thus, we usually work with the number of claims occurred in the exposure period and we observe its realization . The assumption we take is that:

where is the expected value of the yearly number of claims .

Under the GLM assumptions, we obtain

That means that we can model as response variables in a GLM with Poisson response in which the linear predictor depends on an offset additive term .

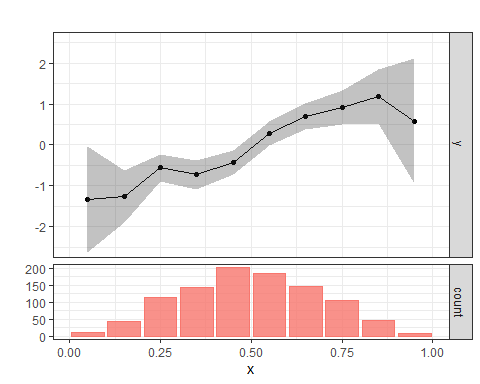
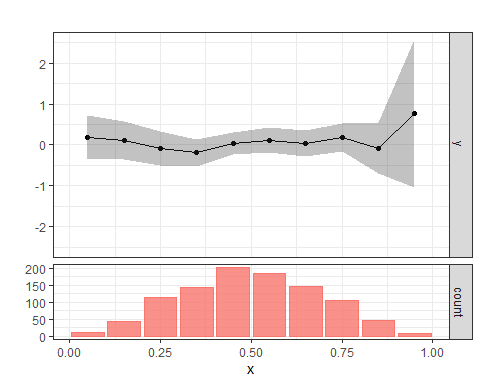
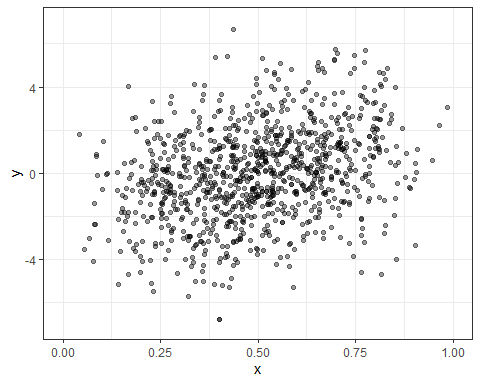
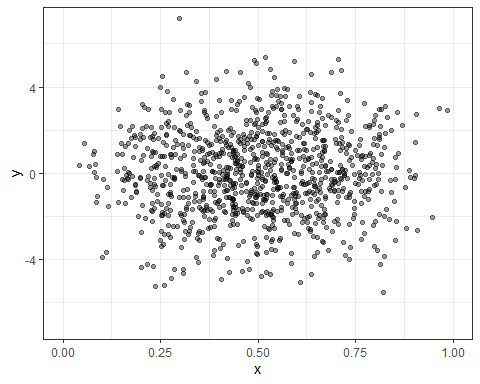
If , the model structure can be expressed as:

Thus, the term can be seen as the multiplicative relativity corresponding to the variable . However, in this case the relativity doesn’t multiply directly the probability of success , but it multiplies the odds of success .

#### 2.1.1.6 Variable selection

One of the most challenging aspects of GLM fitting is selecting the variables and their effects by looking to observed data. In practice, we usually have many explanatory variables available but only some of them are relevant for the prediction of the response variable. Adding useless variables to the model increases the variance of the estimators of the coefficients and then the variance of the predictions . On the other hand, being too frugal with explanatory variables could lead to wasting part of the predictive power of the available explanatory variables.

One useful tool we have to understand if an explanatory variable is relevant or not is plotting the points , as we did in 2.2, 2.6 and 2.10. If there are too many observations and the plot is not easily readable, it is possible to group the points by modalities and show for each group the average of and a confidence interval that gives an idea on the dispersion of the observations around the average. If is a continuous variable with too many modalities, it is possible to group them into buckets. Showing the average of for groups of is particularly useful for Binomial and Poisson data, where the fact that can present few different values compromises the plot readability. An example is reported in figure 2.14. The top-left and bottom-left panels represent a case in which and are not related, while the top-right and bottom-right panels represent a case of positive correlation. From the ungrouped plot in the top-right panel the effect is not clear, while from bottom-right panel it is evident.



If we are dealing with a multivariate model where we already inserted the variables and we want to evaluate the additional information brought by , it is possible to look at the plot , where are the residuals of the model without the variable :

If the plot shows a clear trend, we will add the variable to the model, otherwise we won’t.

The choice of adding or not a variable in the model can be supported by *hypothesis testing*. Given a GLM with coefficients , it is possible to test if a group of coefficients is equal to zero. Formally the hypotheses tested are:

If the hypothesis is accepted, the variables can be removed from the model; if the hypothesis is rejected, at least some of the variables should be kept in the model.

If we want to test if a quantitative variable has a significant effect, we can conduct the test on the single coefficient . If we want to test if a qualitative variable with dummy encoding is significant, we can conduct the test on the coefficients . For qualitative variables it is also possible to conduct a test for each level ; in this case we would test one by one if each level has an effect that is significantly different form the base level effect.

To conduct the test it is possible to use several statistics. One of them is the test based on *log likelihood ratio*. Let’s indicate with the estimated coefficients from the model without any constraint and with the estimated coefficients with the constraints defined by . As the space belongs to is a subset of the space belongs to, it results:

and then:

The basic idea is that, if the variables have a significant effect, will be much higher than and we will reject the hypothesis , while if the variables have not a significant effect, will be more or less the same as and we will accept the hypothesis .

To perform the test, the quantity usually employed is the following:

If we indicate , it is possible to demonstrate that, under the hypothesis , has approximately chi-squared distribution with degrees of freedom:

Therefore, with a significance level we will reject when , where is the quantile of order of the distribution .

The same approach can be used in general for testing hypotheses that can be expressed as , where is a matrix and . This is particularly useful in qualitative variables to test if some of the levels have the same effect and can be then unified. For example, if and are two dummy variables that describe two levels of the same quantitative variable, we can perform the test in order to decide whether unifying the two levels is suitable or not.

Anyway, selecting the variables by performing hypotheses testing have some drawbacks. First of all, conducting a lot of test produce the multiple test problem. Let’s consider the case in which we conduct a test of the kind with a significance level on many variables that have no effect on the response. On average, although the null hypotheses are always true, we will reject them once every 20 tests. That means that if we have available 100 variables, we will randomly select 5 of them, falling in *overfitting*. To fix this problem, it is possible to use the Bonferroni correction, that consist in dividing by the number of test conducted to define the rejection region. But this could be a too restrictive correction that could lead to discard from the model some useful variables, falling in *underfitting*.

Moreover, hypothesis testing aim is finding whether data supports or not an hypothesis. If the aim of the model is prediction, basing variable selection on hypotheses testing could lead to a sub-optimal model.

Another method for variable selection is comparing the models by computing *information criteria*. Two information criteria commonly used are the *Akaike Information Criterion* (AIC) and the *Bayesian Information Criterion* (BIC):

The aim of these statistics is to penalize the likelihood by adding a component that measures the complexity of the model. Among the models considered, the optimal model will be the one that minimizes the information criterion. Thus, if two models have the same likelihood, the optimal model will be the one with less parameters. If the model with more parameter has a higher likelihood, it will be chosen only if that increase in likelihood compensated the increase in complexity.

Another way to compute model predictive performance is by randomly splitting the available dataset into a *training set* (or learning set) and a *test set* , with labeling the dataset and . With this split, we can fit the model on only the training set and assess its performance on the test set by computing the deviance , where is the vector of the coefficients estimated on the training set and is the vector of the observed response variables in the test set. This way it is possible to choose the best model as the one that minimize the deviance in the test set and then fitting it with the whole dataset .

A limit of the train-test approach is that it could bring to overfitting in the test set. Indeed, in particular if the dataset is small, it is possible that a specific set of variables minimizes the deviance on the test set just by chance. To prevent this, it is possible to conduct a *K-fold cross validation*. This consists in randomly partitioning the dataset into subsets and, for each subset , performing a train-test procedure keeping as a training set and as a test set. For each we can compute the testing deviance . We can then compute the average deviance within the subset as:

Thus, the best model will be the one that minimize .

The higher is, the less subjected to randomness is. However, the higher is, the more computationally expensive the procedure is. A common choice for is . If we choose , the procedure is also called *leave-one-out cross validation*.

#### 2.1.1.7 Scalability and manual fitting

One problem of GLM is that the variables selection process is not so easily scalable. Indeed, if we consider explanatory variables, even without taking into account interactions and quantitative variables transformation, there are possible models that can be obtained by choosing a subset of those variables. As increases, building all these models and choosing the optimal one becomes unfeasible.

One strategy to reduce the time consumption is to use a *stepwise procedure*. First of all we must choose a criterion to compare the models, such as the AIC. Then we have to choose a starting model that consider the variables subset . It is possible to compare this model with all the models that can be obtained by removing one variable from or adding to one that is not included in it. From all these models we can compute the AIC and we will choose as the set of variables that minimize the AIC. If none of the considered model has an AIC lower to the one obtained with the variables , the procedure ends and our final set of variables is . Otherwise, we will repeat the step with . The procedure can be iteratively repeated until we obtain a subset of variables that can’t be improved by removing or adding a variable. The model with the variables will be our chosen one.

This procedure is much faster than computing all the models, but it is still not so scalable for large . Moreover, in this procedure we are not taking into account the interactions and the possible transformations for the quantitative variables. This can be achieved by slightly modifying the algorithm, but it would further increase the complexity and the time consumption. Another option is starting from the result of the stepwise regression and manually choosing interactions and quantitative variables transformations by looking at plots as described in 2.1.1.4.

One characteristic of GLM is that the variables effects can be easily interpreted and the variable selection process is for large part manual. This aspect can be problematic if there are a lot of explanatory variables, but it brings some important benefits too. Indeed, the actuary can take choices on variable selection not only based on observed data, but also on his domain knowledge. For instance, the choice of selecting or not an explanatory variable can be guided also on its interpretation: if the observed effect makes sense, it can be added to the model even if it is not statistically significant and it doesn’t decrease the AIC, and, on the opposite, if the observed effect is not reasonable, the actuary can choose not to insert the variable in the model even if its effect is supported by the data. So, in the actuarial practice, in the GLM fitting process there is always a subjective component that impacts on the final result. For these reason it is important for the actuary to have a deep knowledge on the phenomenon he is modeling.

An aspect that facilitates the model building and reduces the complexity of the process, even if there are many variables, is that usually the models are not built from scratch. Actuarial models are usually updated once a year, so it is also possible to start the new model by fitting on new data the final model from the year before. As the models are usually built with policies data from more than one year, the new dataset partially overlaps with the one from the previous year. However, the overlapping observations aren’t identical: the new dataset will have the new settlement information and some new explanatory variables. Anyway, if the actuary is familiar with the effects of the variables in the previous models, he already knows which will probably be relevant and which don’t even deserve too much attention.

### 2.1.2 GAM

In section 2.1.1.4 we have seen that sometimes quantitative variables have not a linear effect on the linear predictor. In GLM it is possible to deal with non-linear effects by adding polynomial or split-wise terms. GAMs are models based on GLM that introduce a more flexible way to deal with quantitative variables with non-linear effects.

#### 2.1.2.1 Model assumptions

In GAM, the assumptions are the same of GLM, stated in 2.1.1.2, with the following advancement in the linear predictor:

where

* is the vector of the variables with a linear effect as described in GLM, that also include a term that represents the intercept;
* is the vector of the linear coefficients as described in GLM;
* are the quantitative variables with a non linear effect;
* are continuous functions with continuous first and second derivatives.

The functions introduce the possibility to model non-linear effects of the variables .

#### 2.1.2.2 Cubic splines

A class of functions commonly used for modeling is the class of the *cubic splines*.

Definition 2.3: (Cubic Splines) Let consider real numbers , called , and a function such that:

where, for , . For the last index , is extended to .

is a if it satisfies the following conditions in the internal knots $$

The constraints for make it a continuous functions with first and second derivatives continuous in . It is possible to extend the cubic spline to an interval with linear extensions on and . The so built function is called *natural cubic spline*.

In definition 2.3 we introduced 4 parameters () for each of the intervals , . So, we have parameters. In equation (??) we introduced constraints for each of the knot in . So the free parameters become . The linear extension on and corresponds to the constraint on , thus, and . Adding these two constraints, we get that the natural cubic spline has degrees of freedom.

With an approach similar to the one we used in 2.1.1.4 for split-wise effects in GLM, we can represent a cubic spline by the functions , .

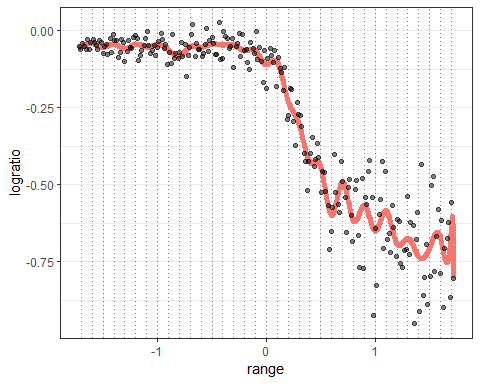
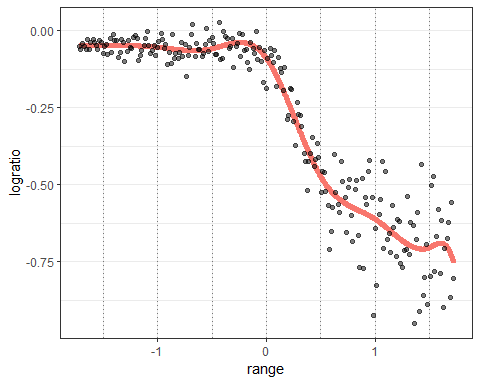
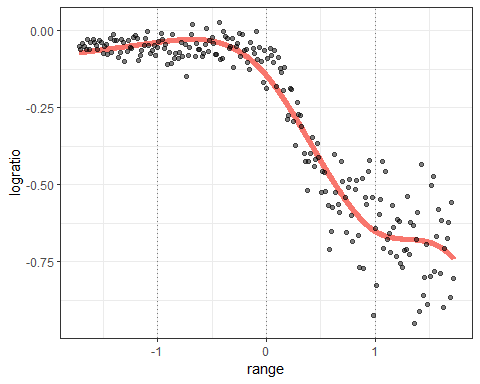
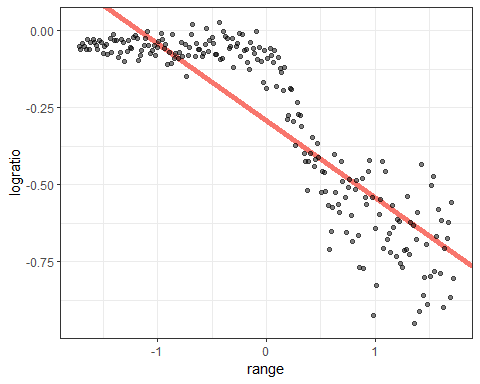
The expression:

gives a natural cubic spline.

The two side constraints ensure that we have a smooth linear extension above . The expression (??) presents parameters, thus, with the 2 side constraints, there are degrees of freedom. From this expression it is easy to show that the natural cubic splines over the interval with knots constitute a -dimensional vectorial space.

#### 2.1.2.3 Smoothing

If we try to fit a cubic spline on data, we find that by increasing the number of knots, the function tends to overfit data. This is due to the fact that, increasing the number of knots, we are increasing the number of parameters of the model and, thus, the variance of the parameters estimators increases. We can see this effect in figure 2.18.



In the top-left panel, the model fitted is just linear and it clearly doesn’t follow the path made by the observed points. In the bottom-right panel, the models fitted is a spline with 35 knots and it is clearly too wiggly compared to the path made by the observed points.

A measure of the wiggliness of a function is given by the integral of its squared second derivative:

In figure 2.22 some examples of functions and their squared second derivatives are represented. In the top panels, four functions of increasing wiggliness are represented and, in the bottom panels, we can see their four squared second derivatives.

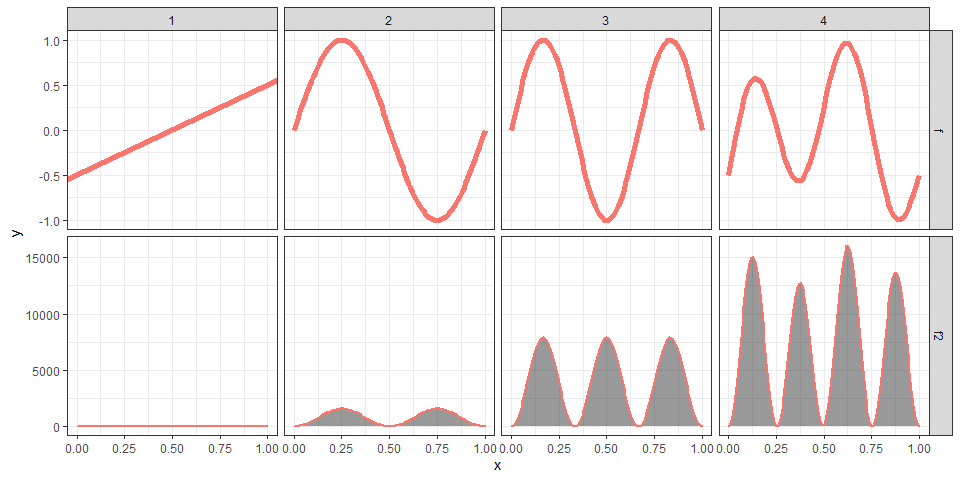


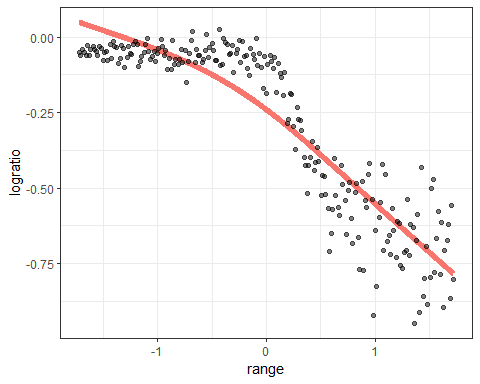
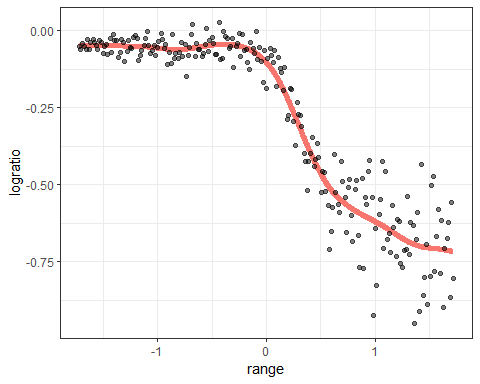
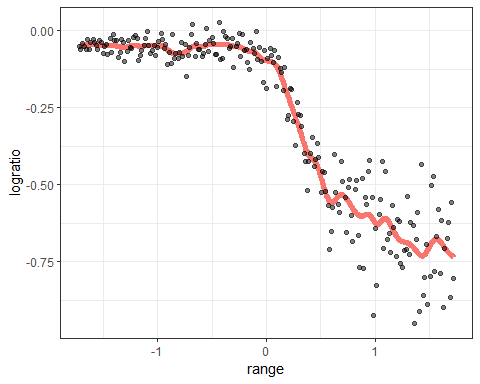
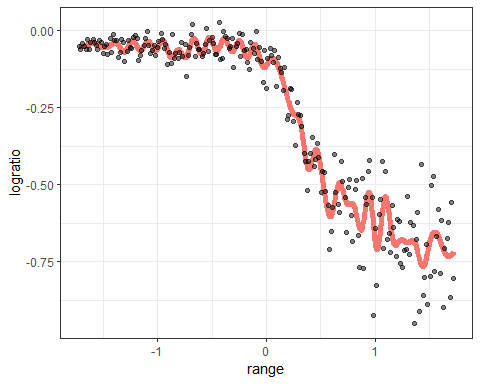
Figure 2.22: for functions with different wiggliness.

As we can see from the plots, if a function is linear, is null and, as the curvature of the function increases, increases. For this reason, can also be seen as a measure of how much differs from a linear function.

What is done in GAM to contain this wiggliness and to prevent overfitting is taking into account a penalization based on . If we have a regression problem with one explanatory variable , this can be achieved by considering the minimization problem in (??) and adding a penalization term to the deviance:

$$\begin{equation}
\label{eq:gam-est-deviance}
\hat{f} = \argmin\_{f}{\left\{D(f, \boldsymbol{y}) + \lambda \int\_{a}^{b}{\left( f''(x) \right)^2 dx}\right\}}
\end{equation}$$

In the notation of the formula (??) we used to indicate all the parameters of the natural cubic spline. The hyper-parameter is called *smoothing parameter*. It measures how much we want to penalize wiggliness. If we choose we won’t penalize for and the optimization problem corresponds to the maximum likelihood. The higher is, the more penalization for we introduce and the smoother the estimate will be. If , we will introduce an infinite penalization for wiggliness, so the result will have , thus will be linear. An example with different levels of can be seen in figure 2.23.



All the models has been fitted with knots. As we can see, with we are clearly overfitting observations, while with we are clearly underfitting them.

As soon as the number of knots is large enough, the exact number and the positioning of the knots is not important. If the function is flexible enough, the tuning of the curve is done by just tuning . It is possible to use as many knots as the determinations of are, but it could be too computationally expensive. A possible choice is to fix a number of knots and positioning on empirical quantiles of the observed or equally spaced in the range of . In our example, knots seems to be large enough.

In general, if we have quantitative explanatory variables that we want to fit with splines, the formula (??) becomes:

$$\begin{equation}
\label{eq:gam-est-deviance-multi}
\boldsymbol{\hat{f}} = \argmin\_{\boldsymbol{f}}
{\left\{
D(\boldsymbol{f}, \boldsymbol{y})
+ \sum\_{l=1}^{q}{
\lambda\_l \int\_{a\_l}^{b\_l}{\left( f\_l''(x\_l) \right)^2 dx}
}
\right\}
}
\end{equation}$$

where .

In this context, we have to tune a vector of smoothing parameters:

#### 2.1.2.4 Choice of the smoothing parameters

As described, the selection of the effect consists just in finding the optimal parameter . The technique commonly used in machine learning for hyper-parameter tuning is the *Cross Validation*.

For a set of values of we can perform a K-fold cross validation as described in 2.1.1.6 and, for each , we can compute the average test deviance:

At this point, we will choose the hyper-parameter vector that minimizes the cross-validation deviance:

$$
\hat{\boldsymbol{\lambda}} = \argmin\_{\boldsymbol{\lambda}\in\Lambda}{D^{CV(K)}\_{\boldsymbol{\lambda}}}
$$

In particular if is big, this procedure can be too computationally expensive. Two alternatives are the *Generalized Cross Validation* (GCV) and the *Un-Biased Risk Estimator* (UBRE). The idea behind these approaches is to estimate the test set deviance by just computing the training set deviance and applying to it a correction that penalizes for the complexity of the model, as it is done in AIC and BIC, described in section 2.1.1.6.

GCV and UBRE formulas are based on the fact that GAMs with identity link and Normal response are *linear smoothers*, i.e.  can be expressed as:

where is a matrix that depends on the design matrix and the hyper-parameters. is called *smoothing matrix*. In general, if the link is not identity and the response is not Normal, GAM is not a linear smoother, but the formula (??) still holds up with a local approximation.

It can be proven that the trace of measures the flexibility of the function. Indeed, in Linear Model and , that is the number of *degrees of freedom* of the model. For this reason is also called number of *effective degrees of freedom* of the model. In GAM, as the smoothing parameter increases, the flexibility of the model decreases and decreases.

For linear smoothers it can be proved that the *leave one out cross validation* is:

where is the th element of the diagonal of the smoothing matrix .

This formula is particularly convenient because it requires just the fit with the whole dataset and not one fit for each fold. Indeed, in the formula the quantities that must be computed are and , that both depend on the model fitted with the whole dataset.

This formula can be further simplified by replacing the values with their average . This way we get the GCV.

As for K-fold cross validation we can compute the GCV for different values of and choose the value that minimizes the GCV.

The formula (??) can be expressed as:

The expression (??) generalizes the GCV to all the GAMs, also in the case in which the link is not identity and the response is not Normal.

To select the best value of , we can perform a similar procedure using the UBRE in place of the GCV. The UBRE is defined as:

Using the UBRE instead the GCV is preferred when is known, such as in Poisson regression case, in which .

#### 2.1.2.5 Why cubic splines

In section 2.1.2.2 we said that *cubic splines* are commonly used for modeling in GAMs. This choice comes from the following theorem.

Theorem 2.1: (Spline property) Given the knots and the values , for any and , only one natural cubic spline , such that , exists.

Moreover, given a function two times differentiable with continuity, such that , then, for any and , it results

One consequence of this theorem is that, within all the continuous function with continuous first and second derivatives, the one that minimize the optimization problem (??) is always a natural cubic splines.

Let’s consider the determinations of the variable . If is a continuous function with continuous first and second derivatives, for the theorem 2.1, there exists only one natural cubic spline such that . As, , it results that . As and , it results that, for any given :

For this reason, if the aim of the model estimation is to minimize (??), for choosing we can just consider the class of the natural cubic splines.

#### 2.1.2.6 Other basis

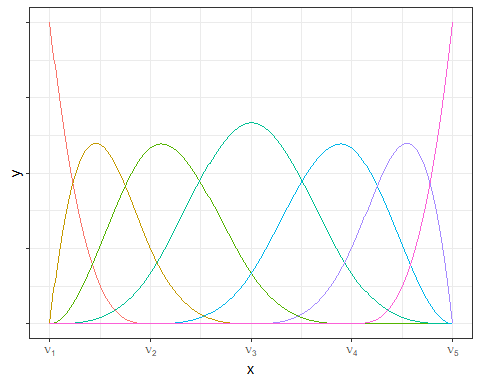
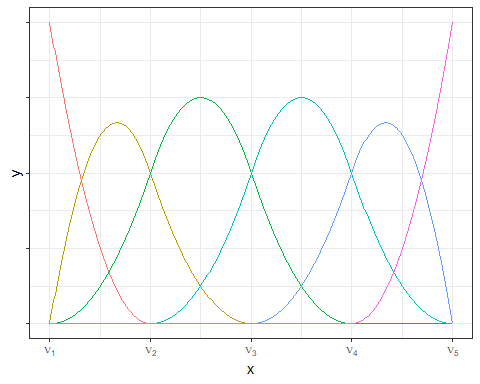
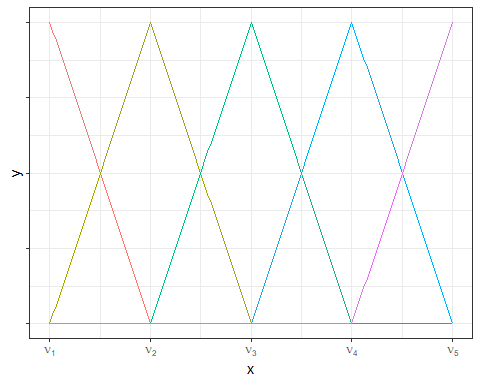
In section 2.1.2.2 we said that natural cubic splines on knots constitute a -dimensional vector space and that a possible basis decomposition is (??).

Another usual basis is the *B-spline* basis. B-splines are functions defined recursively as follows.

Definition 2.4: (B-splines) For :

For and :

In figure 2.27 some B-splines of different degrees are represented.



It can be proven that 2.4 defines functions that, given , are splines of degree . Moreover, constitute a basis of the vector space of the splines of degree on the knots . In particular, constitute a basis of the vector space of the splines of degree . Therefore, if is third degree spline on the knots , it can be expressed as a linear combination of :

B-splines are preferred compared to truncated polynomial as they are less correlated and lead to more stable and less computationally expensive estimates.

#### 2.1.2.7 GAM extensions

As in GLM, in GAM we can consider interactions between a variable with non-linear effect and another variable . This can be achieved by adding to the linear predictor a term such as:

This is particularly useful when is a binary variable and we want to fit two different curves for in the case and in the case .

If we want to consider a more complex interaction between two quantitative variables with non-linear effect, GAM can be extended by considering non-parametrical interactions and modeling them with two-dimensional splines, such as:

In this case, instead of fitting a curve on , we will fit a flexible surface on . This approach can be adopted also for modeling geographical data, in which are coordinates on the map, such as longitude and latitude.

#### 2.1.2.8 Some considerations on GAM

As we have seen in this chapter, GAMs are flexible tools for fitting quantitative variables with non-linear effects.

One big advantage of GAM is that they are based on the same assumptions of GLM, except for the non-linearity of the components . The connection with GLM leads to highly interpretable results. Indeed, as we do in GLM, in GAM we can easily observe the marginal effect of a variable on the response just by plotting the graph , while the interactions between variables are added manually so we have a full control of them.

Another big advantage is that GAM not only provides a flexible tool that produces interpretable results, but this tool works almost automatically. Indeed, while in GLM, when we have to fit a non-linear effect to a quantitative variable , we have to perform a manual process of wise splitting and polynomial fitting on the range of , in GAM we do not have to explicitly specify the shape of and everything is done by the algorithm.

This higher flexibility and automation comes at the cost of introducing more complexity in the model. Indeed, in a GAM there are much more parameters than in a GLM and this leads to a more computationally expensive fitting. Anyway, this higher complexity produces an higher machine time but significantly reduces the human time that in GLM would be needed for manual fitting.

### 2.1.3 Shrinkage estimators for GLM

In this section we are going to present the shrinkage estimators, that are a class of estimators particularly useful in GLM when there are many explanatory variables. All the models presented in this chapter are actually GLM. The advancements consist in the way the parameters are estimated.

#### 2.1.3.1 The Bias-Variance Trade Off

One of the property of GLM with linear link and Normal response is that the maximum likelihood estimator is unbiased, that is .[[4]](#footnote-129) However, the bias is not the only relevant aspect of an estimator. If we consider the *Mean Squared Error* (MSE) of an estimator , we get the following result:

The idea behind shrinkage estimators is to add an amount of smart bias to in order to reduce its variance.

The trade-off between Bias and Variance described by equation (??) can be interpreted in relation to the complexity of the model. Figure 2.30 shows how the prediction error changes by increasing the complexity of the model. In linear smoothers, the complexity of the model can be measured as the effective degrees of freedom (see section 2.1.2.4), that in the case of the GLM with maximum likelihood estimators is just the number of parameters . Usually, in machine learning models, the complexity of the model is determined by tuning one or more hyper-parameters. In the case of GAM, for example, by increasing we obtain a less complex model, while by decreasing it we obtain a more complex model.

A model not complex enough produces estimators with low variance but high bias, that leads to underfitting. A too complex model produces estimators with low bias but high variance, that leads to overfitting.

As with maximum likelihood estimators for GLM the complexity increases with , they are not suitable when is large (high dimensionality).

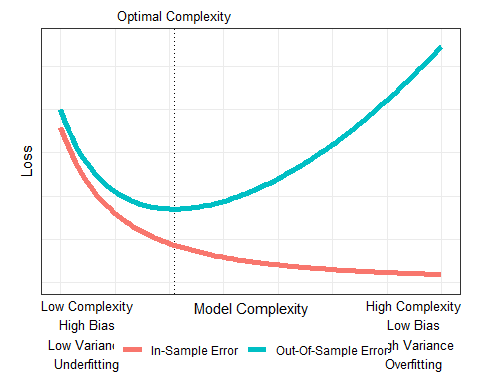


Figure 2.30: The Bias-Variance trade off. The horizontal axis shows the increase in complexity and the vertical axis shows the effect on in-sample error (training set) and out-of-sample error (test set). The error can be measured as the Mean Squared Error (MSE) or in general as the Deviance.

#### 2.1.3.2 Ridge Regression

One shrinkage estimator for GLM is the *Ridge Regression*. In Ridge Regression, the decrease in variance of is achieved by considering the optimization problem (??) and adding to the Deviance a penalization term that depends on the magnitude of :

$$\begin{equation}
\label{eq:ridge-est-deviance}
\hat{\boldsymbol{\beta}} = \argmin\_{\boldsymbol{\beta}\in\mathbb{R}^{p+1}}{\left\{D(\boldsymbol{\beta}, \boldsymbol{y}) + \lambda \|\boldsymbol{\beta}\_{\setminus0}\|\_2^2\right\}}
\end{equation}$$

where:

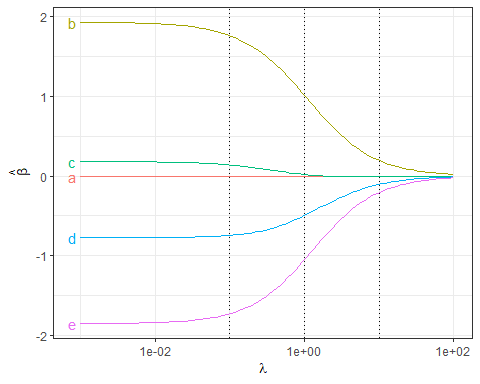
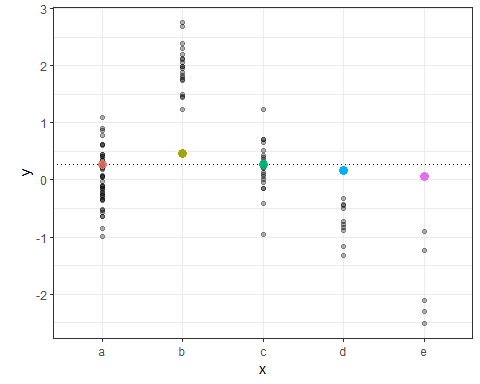
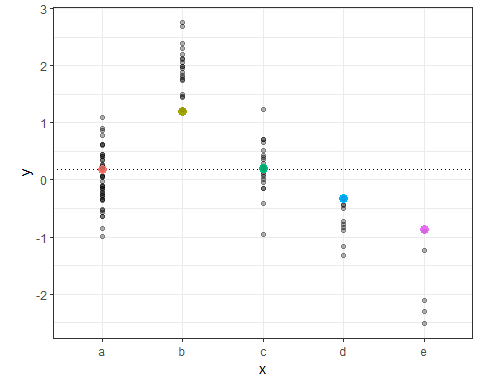
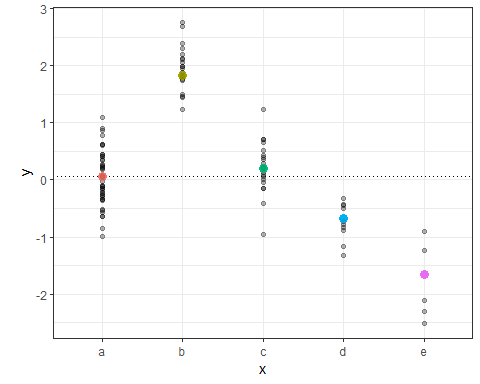
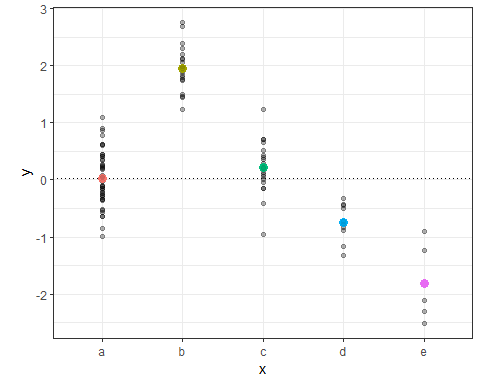
* is the set of the GLM coefficients except for the intercept ;
* is the norm, i.e. ;
* is an hyper-parameter that controls the penalization.

The term produces a penalization for high values of . This penalization leads to a shrinkage of the coefficients. The exclusion of from the penalization term is intended to prevent the introduction of a bias towards in the intercept. As the magnitude of depends on the values of , it is preferred to standardize all the explanatory variables to avoid distorting effects due to the unit of measure of the explanatory variables.

As for GAM, the hyper-parameter determines the amount of penalization given to high values of the coefficients . If , the optimization problem (??) corresponds to the maximum likelihood. As increases, the optimal coefficients tends to shrink towards . In the limit case, if , the optimal coefficients are all equal to , except for , that is equal to , so the estimated model corresponds to the trivial model with only the intercept .

An example of the effect of is reported in figure 2.31. The data represented has been simulated from a GLM with idenity link and Normal response. As we can see, if , all the estimated responses correspond to the average of the response on that group . As increases, the estimated responses move towards the global average . The speed of convergence to depends on how much the group average differs from the global average and on the number of observation of the group: if in a group there are many observations, we have a lot of information on that group, the group average have a small variance and it is a reliable estimate for , while if there are few observations, the group average have a high variance and it is not very reliable.

The optimal point for can be obtained through a Cross Validation as seen in section 2.1.2.4.



The case of GLM with identity link and Normal distribution is particularly convenient for the interpretation, because, in this case, the optimization problem (??) has an explicit solution. To make the results more interpretable, let’s assume that the response variables have and that the explanatory variables are centered on , i.e. . That means that and the model is . With these assumptions, we obtain:

where:

* ;
* is the design matrix without the first column of s, as the intercept is excluded from the model;
* is the identity matrix with dimension .

From the formula (??) we can see that, while the maximum likelihood estimator is unbiased, if the Ridge estimator is biased.

Moreover, from the formula (??) we find that, even if there is multicollinearity and is not invertible, the Ridge estimator is computable. This aspect is particularly interesting when .

If we assume that the explanatory variables are independent and standardized, i.e. , we can further simplify the expression (??) to:

that results in:

This result means that, if the explanatory variables are independent, the Ridge penalization shrinks all the estimated coefficients by a factor of and reduce their variance by a factor of .

From formula (??) we also find that, as in GAM, if the link is identity and the response is Normal, the Ridge regression is a linear smoother. In this case the smoothing matrix is:

In the case of independent explanatory variables we get:

and then . This result means that, by increasing , the effective number of degrees of fredom decreases.

#### 2.1.3.3 LASSO Regression

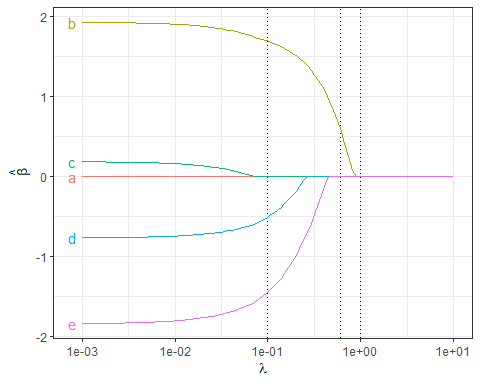
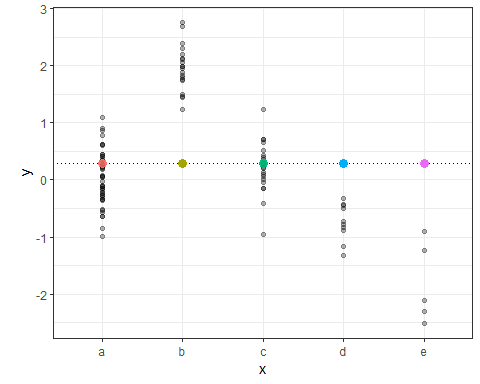
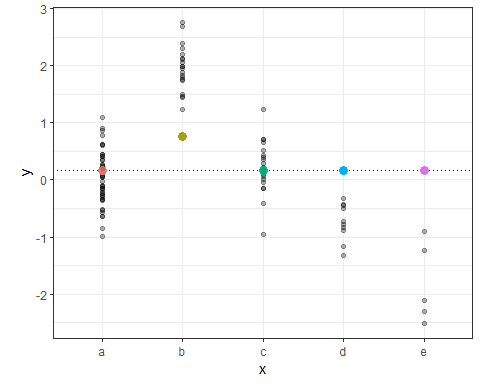
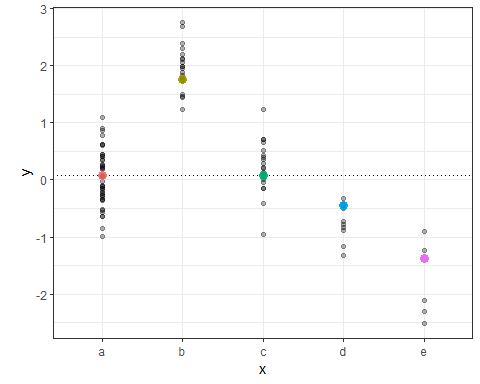
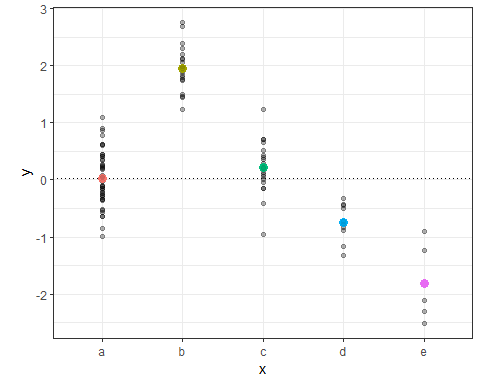
Another shrinkage estimator for GLM is the *LASSO* (Least Absolute Shrinkage and Selection Operator). LASSO is based on the same idea of Ridge Regression, but, instead of considering a penalization based on norm, it considers a penalization based on norm:

$$\begin{equation}
\label{eq:lasso-est-deviance}
\hat{\boldsymbol{\beta}} = \argmin\_{\boldsymbol{\beta}\in\mathbb{R}^{p+1}}{\left\{D(\boldsymbol{\beta}, \boldsymbol{y}) + \lambda \|\boldsymbol{\beta}\_{\setminus0}\|\_1\right\}}
\end{equation}$$

where:

* is the set of the GLM coefficients except for the intercept ;
* is the norm, i.e. ;
* is an hyper-parameter that controls the penalization.

An example of the effect of the penalization for different values of is shown in figure 2.36. The simulated dataset is the same used for the Ridge example in figure 2.31. As we can see from the plots, the substantial difference between Ridge and LASSO is that in LASSO from a certain value of the coefficients are shrunk exactly to . While in Ridge that is just a limit property, in LASSO, for each coefficient there is a level of the penalization parameter such that for , the estimated coefficient is forced to exactly .



This property of the LASSO Regression can be derived from a different representation of the optimization problems (??) and (??). It can be proven that in general, considering a penalization given by the norm, the unconstrained optimization problem:

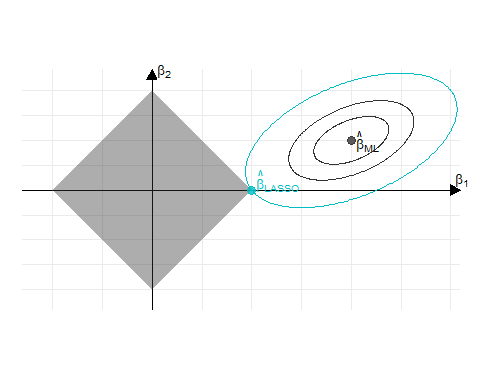
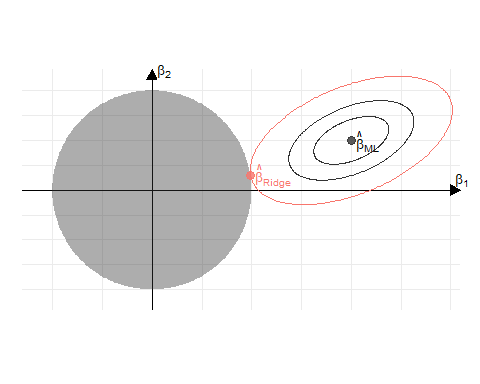
$$\begin{equation}
\label{eq:ld-est-deviance}
\hat{\boldsymbol{\beta}} = \argmin\_{\boldsymbol{\beta}\in\mathbb{R}^{p+1}}{\left\{D(\boldsymbol{\beta}, \boldsymbol{y}) + \lambda \|\boldsymbol{\beta}\_{\setminus0}\|\_d^d\right\}}
\end{equation}$$

is equivalent to the constrained optimization problem:

$$\begin{equation}
\label{eq:lp-est-deviance-constr}
\hat{\boldsymbol{\beta}} = \argmin\_{\boldsymbol{\beta}\in\mathbb{R}^{p+1}:\ \|\boldsymbol{\beta}\_{\setminus0}\|\_d^d \le s\_{\lambda}}{ \left\{ D(\boldsymbol{\beta}, \boldsymbol{y}) \right\} }
\end{equation}$$

where is a quantity that depends on .

The representation (??) provides a useful geometric interpretation of the optimization problem. Figure 2.41 shows the visual representation of the optimization problem (??) in the Ridge case () and in the LASSO case (). The axes represent the component of . The point represents the maximum likelihood estimator for , that is the optimal point for the Deviance without any constraints. The ellipses around represent the contour lines of . In the Normal case with identity link they are concentric ellipses centered in . The grey area around the axes intersection represents the feasibility region determined by . This area, in the Ridge case corresponds to the circle , while in the LASSO case corresponds to the square . The sharpness of the LASSO feasibility region imply that the optimal point could fall into one of the corners of the square leading one of the coefficients to be exactly equal to . In general, if there are more than two explanatory variables, the LASSO feasibility region is an hyper-cube and the LASSO attains solutions with many coefficients exactly equal to .



The fact that in LASSO Regression the estimated coefficients can be exactly equal to is a precious benefit. Indeed, LASSO Regression performs a feature selection removing the coefficients that are not relevant for predicting the response. The LASSO fitting is much more efficient than the other procedures we have seen for feature selection in GLM such as the stepwise selection based on AIC or other criteria (section 2.1.1.6) and it is better scalable for datasets with many variables. It is also possible to use the variables selected by the LASSO regression and giving them as input for a maximum likelihood fitting. If in the dataset there are many variable but only few of them are relevant for the response, this procedure can returns better predictions than just LASSO regression.

#### 2.1.3.4 Elastic Net

Compared to the Ridge Regression, the LASSO Regression has the benefit of performing a feature selection by forcing many coefficients to . Anyway this doesn’t necessarily mean that the LASSO produces estimates that always outperform Ridge estimates. It strongly depends on the case. In general, if in a dataset only few of the explanatory variables have an effect on the response, the LASSO will produce better estimates, but, if all the variables bring a small amount of information on the response, the the LASSO will suppress some of this information, while the Ridge will catch it. The problem is that in practice, when we have a real dataset, we do not know in which case we are. The *Elastic Net* is a generalization of the Ridge and the LASSO that provides a solution for this kind of problem.

The Elastic Net consists in a penalized Deviance optimization problem in which the penalization term is a mixture of the Ridge penalization and the LASSO penalization:

$$\begin{equation}
\label{eq:elastic-net-est-deviance}
\hat{\boldsymbol{\beta}} = \argmin\_{\boldsymbol{\beta}\in\mathbb{R}^{p+1}}{\left\{
D(\boldsymbol{\beta}, \boldsymbol{y}) +
\lambda \left(
\sum\_{j=1}^p{\alpha |\beta\_j| + (1 - \alpha) |\beta\_j|^2}
\right)
\right\}}
\end{equation}$$

where is an hyper-parameter that weights the two penalization components.

Looking to the equation (??) it is clear that, if , the Elastic Net corresponds to the Ridge Regression, while, if , the Elastic Net corresponds to the LASSO regression. If , the result will be a compromise between the two.

The hyper-parameter can be estimated together with in a Cross Validation procedure (see section 2.1.1.6). If the data suggests that many variables are useful for predicting the response, the estimated hyper-parameter will be close to , while if the data suggests that only few variables are useful, the estimated hyper-parameter will be close to .

#### 2.1.3.5 Some considerations on Shrinkage Estimators

As we have said, shrinkage estimators are particularly useful when the number of parameters in the model is large (high dimensionality) and the maximum likelihood estimator have a high variance. This could happen when we have some qualitative variables with many modalities, as for example the make and the model of the insured vehicle in car insurance. As we have seen with the examples represented in figure 2.31 and 2.36, the penalization will shrink more the coefficients in the groups with few observations and only the groups with an average response significantly different from the global average response will emerge. This procedure is more efficient and more robust to overfitting than grouping modalities based on hypothesis testing or other criteria.

If we are in a case in which we already performed a satisfying feature selection and we only want to shrink some of the coefficients we can apply the penalization only to them. For example, in the case of the variables make and model, we can consider an Elastic Net penalization only to the coefficients corresponding to the modalities of those variables. This technique is useful also in the case we have explanatory variables with distribution highly unbalanced. For example, if we consider the variable “number of claims experienced in the previous year”, we will have most of the observation in the modality corresponding to claims, very few in the modality corresponding to claim and almost nobody with or more claims. If we are fitting a model for the claim frequency and we consider the variable that indicates whether the policyholder experienced one or more claims in the previous year or not , it is likely that the maximum likelihood estimator for the coefficient will be greater than , as the policyholders that experienced claims are usually more inclined to experience more of them in the future. However, given that in there are just few observation with it is possible that the coefficient is not significantly different from . If our only tool is the maximum likelihood estimator, we have to choose whether to insert the variable in the model or not. If we don’t consider it we are probably discard some potentially useful information, while if we consider it and we estimate its coefficient with maximum likelihood we risk to overfit the data. With shrinkage estimators we can choose to insert the variable in the model and fitting it with a penalization. This way we exploit that information but we prevent overfitting.

Another interesting observation about GLM fitting is that in the practice, when maximum likelihood estimates are adopted, what is done is not just fitting a model with all the variables, but a feature selection is conducted. The fact that a variable is inserted in the model depends on whether the feature selection procedure selects that variable or not. That corresponds to estimating the coefficient with an estimator that is equal to the maximum likelihood estimator when the coefficient pass a specific criterion and is equal to when doesn’t pass that criterion. This criterion could be something objective as for example the decrease in AIC and the significance in a hypothesis testing on , and also something more subjective based also on the domain knowledge on the person that is conducting the modeling. Actually, this feature selection procedure introduce a bias towards of the coefficient that reduce its variance as all the coefficients not relevant for prediction are set equal to and only the relevant ones are fitted with maximum likelihood. If we consider a binary variable , in all the procedures commonly used for feature selection, the probability that passes the procedure or not depends on how strong the effect is and how many observation there are in the classes and .

We also mantion that the Shrinkage Estimators can be used in synergy with GAMs. Indeed it is possible to fit to the quantitative variables a cubic spline with a GAM penalization based on the second derivative of the spline and to the qualitative variables a shrinkage estimator with an Elastic Net penalization.

$$\begin{equation}
\label{eq:gam-en-est-deviance-multi}
\boldsymbol{\hat{f}} = \argmin\_{\boldsymbol{f}}
{\left\{
D(\boldsymbol{f}, \boldsymbol{y})
+ \sum\_{l=1}^{q}{
\lambda\_l \int\_{a\_l}^{b\_l}{\left( f\_l''(x\_l) \right)^2 dx}
}
+ \lambda\_{EN} \left(
\sum\_{j=1}^p{\alpha |\beta\_j| + (1 - \alpha) |\beta\_j|^2}
\right)
\right\}
}
\end{equation}$$

where the coefficients considered in the Elastic Net penalization are only the ones corresponding to qualitatives variables.

### 2.1.4 Bayesian GLM

#### 2.1.4.1 Bayesian framework

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.

#### 2.1.4.2 Bayesian estimators for GLM

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.

#### 2.1.4.3 LASSO and Ridge Regression as Bayesian estimators

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.

#### 2.1.4.4 Some considerations on Bayesian estimators

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.

## 2.2 Considerations on models

### 2.2.1 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.

### 2.2.2 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.

## 2.3 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.

# Bibliography

This is my bibliography

1. [ANIA yearly statistical report for motor third party liability](https://www.ania.it/ricerca-avanzata/-/asset_publisher/XIyLeujL9irt/content/id/113283) [↑](#footnote-ref-46)
2. [IAA, About Actuaries](https://www.actuaries.org/iaa/IAA/About_the_IAA/About_Actuaries/About%20Actuaries.aspx) [↑](#footnote-ref-70)
3. Simpson’s paradox, <https://en.wikipedia.org/wiki/Simpson%27s_paradox> [↑](#footnote-ref-86)
4. This property in general is not true for GLM with other links and response distributions, but it is true asymptotically: $\boldsymbol{\tilde{\beta}}^{ML} \xrightarrow{n\to+\infty} \boldsymbol{\beta}$. [↑](#footnote-ref-129)