SCHOOL OF
ECONOMICS &
MANAGEMENT
LISBON

MASTER OF SCIENCE IN

ACTUARIAL SCIENCE

MASTERS FINAL WORK

PROJECT

CALCULATING BEST ESTIMATES IN A GLM FRAMEWORK. FREQUENCY/SEVERITY MODELS VS TOTAL LOSS MODELS.

ALEXANDRE FILIPE CORREIA CAJANA VERMELHO

MAY 2014

ISEG
SINCE 1911
SCHOOL OF
ECONOMICS &
MANAGEMENT
LISBON

Master of Science in
Actuarial Science

Masters Final Work

Project

Calculating Best Estimates in a GLM Framework. Frequency/Severity models VS Total Loss models.

Alexandre Filipe Correia Cajana Vermelho

Supervisor:

João Andrade e Silva

May 2014

Acknowledgments

I would like to start by sending a word of appreciation to Associação Portuguesa de Seguradores for sponsoring this master degree. Also, in a more personal level, to my colleagues Miguel Guimarães, Luís Malcato, Salomé Valente and Vera Moura for giving me all the time and support I needed.

Another word of great appreciation goes to professor João Andrade e Silva for his time, patience, comments, corrections and for the interest he showed in this work. Thank you also to all my other professors and colleagues at ISEG who helped me a lot over the master.

To all my family, specially my mother, my sister, my grandmother and my newborn nephew Gabriel to whom I dedicate this work, I send a word of love. To all my friends and special ones who got very little time from me in these last few months, I'm sorry I took so long.

To my father, who is still the best professor I have ever had.

Contents

Acronyms

APS: Associação Portuguesa de Seguradores

AH: Accidents&Health

BE: Best Estimate

WC: Workers Compensation

CMR: Commerce Multiple Risks

FODP: Fire and other damage in property

HMR: Habitation Multiple Risks

IMR: Industrial Multiple Risks

MI: Motor Insurance

MR: Multiple Risks

Definitions

1. **Non-Life insurance –** Policy agreement between two parties, in wich one of them (the insurer) engages to compensate the other (the policyholder) for a certain unpredictable loss in a fixed time period in exchange of a fee (insurance pemium).
2. **Claim –** Event for wich the policyholder demands finantial compensation from the insurer.
3. **Claim size –** Money paid by the insurer to the policyholder as the result of a claim.
4. **Total claim size of a policy –** Sum of the all the claim sizes made during the fixed time period the policy was valid.
5. **Duration of a policy –** Amount of time a policy is valid.
6. **Annualised exposure –** Fraction of the year the policy was valid, i.e, the duration of the policy measured in years.
7. **Claim frequency –** Number of claims divided by the annualised exposure.
8. **Claim severity –** Total claim size divided by the number of claims, i.e, the average claim size per claim.
9. **Pure premium –** Total claim size divided by the annualised exposure.

Abstract

When using generalized linear models to predict future claim payments, should actuaries use separate frequency/severity models or a single loss cost model? This is the question this paper addresses, covering some theoretical background, testing both alternatives on real data from the Industrial Multiple Risks (IMR) sub-branch and analysing its results. Data was provided by 7 companies operating in Portugal in the years 2010 and 2011, who own a 70% share of the Portuguese IMR market and was collected by Associação Portuguesa de Seguradores (APS).

Keywords: Generalized Linear Models, Tweedie families of distributions, Fire and other damage in property, Multiple Risks, Lift Chart, Best Estimate, Solvency II, over-dispersion in GLMs, Quasi-Poisson distribution, Gamma distribution, offsets

1. Introduction

This work aims to present and compare the results of two different approaches used to calculate estimates for future claim values. Both approaches take advantage of the generalized linear models (GLMs), a set of regression models that has been proved usefull in forecasting, credibility, loss reserving and other actuarial problems, since their first presentation by Nelder & Wedderburn (1972) and their first actuarial illustrations by McCullagh & Nelder (1989).

The usual approach to a tariff/forecasting problem consists in treating claim frequency and claim severity separately, assuming no correlation between these variables. Usually, for a policy or groups of policies, a GLM with a Poisson or Quasi-Poisson distribution is used to fit the claim numbers and a Gamma distribution models the claim severity adequately. This approach is widely used and thoroughly studied by Klugman S., Panjer H. and Willmot G. (2008).

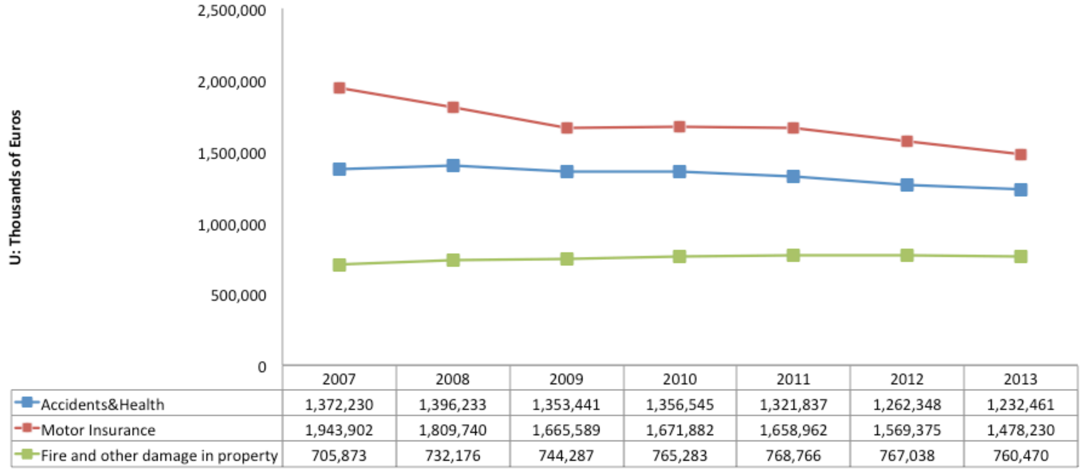
Another approach consists on modeling the total loss of a policy or groups of policies. In this latter case, the Tweedie families of distributions – in view of Tweedie (1984) - are a valid alternative to actuaries. These families of distributions will be presented in chapter 5, but for the scope of this introduction it's enough to know that this approach assumes, in counterpoint with the usual approach, that predictors simultaneously increase or decrease both claim frequency and severity. The value of a predictor is therefore the result of both these effects, making it impossible to have an explaining variable in the model influencing claim frequency and claim severity in different ways, as it might happen when we model these effects separately.

Both alternatives will be tested using real data from the Portuguese Industrial Multiple Risks (IMR) sub-branch, collected by Associação Portuguesa de Seguradores (APS) from seven insurance companies operating in Portugal in the years 2010 and 2011, and whose market-share in the IMR sub-branch reaches up to 70%. The IMR is a sub-branch with little policy exposure (about 40 thousand policies exposed in 2013), where most of the losses are small, but where a single loss can reach several millions of Euros. Our objective is to use the 2010 data to build the models and then compare the predictions both alternatives yield with the real 2011 losses.

In chapter 2 we present a brief study of the Portuguese Non-Life business, focusing on its three major LoB's: The Motor Insurance (MI) LoB, the Fire and Other Damage in Property (FODP) LoB and the Accidents and Health (AH) LoB. In chapter 3 we describe the database used and the reasons for the choice of some explanatory variables. In chapter 4 we give an overview on GLMs and some variations of the model that will be used later, such as the offsets and over-dispersion. In chapter 5 we present the Tweedie families of distributions. In chapter 6 we analyse the fitted models, focusing on variable behaviour and goodness of fit. In chapter 7 we discuss the results both approaches produced in predicting claim amounts and in chapter 8 we draw our conclusions.

2. The Fire and other damage in property LoB and the Multiple Risks branch – a framework in the Portuguese Non-Life Business

Fire and Other Damage in Property (FODP) is currently the third largest LoB in the Portuguese Non-Life insurance industry in terms of written premium production, following the Motor Insurance (MI) and the Accidents and Health (AH) LoBs. The chart below shows the behaviour of the annual written premium of these LoBs over the past 7 years in Portugal:



	2007	2008	2009	2010	2011	2012	2013
🟦 Accidents&Health	1,372,230	1,396,233	1,353,441	1,356,545	1,321,837	1,262,348	1,232,461
🟥 Motor Insurance	1,943,902	1,809,740	1,665,589	1,671,882	1,658,962	1,569,375	1,478,230
🟩 Fire and other damage in property	705,873	732,176	744,287	765,283	768,766	767,038	760,470

Figure 1: Annual Written Premiums in Portugal for MI, AH and FODP. Data collected from the report "Produção Anual de Seguro Direto 2013" compiled by Associação Portuguesa de Seguradores (APS)[1].

Even if lower in value, only the FODP premiums increased in these last 7 years (+7.7%), while the MI and AH premiums decreased (-24%) and (-10.2%) respectively.

We will now dig deeper in these LoB's branches to find different patterns in the premium production. The FODP LoB is composed by several branches that reflect

[1] https://segurdata.apseguradores.pt

potential losses in property, wether in a business or in a household. The following table shows the premium distribution of the LoB among these different branches:

	2007	2008	2009	2010	2011	2012	2013
Fire and Other Damage in Property	**705 873**	**732 176**	**744 287**	**765 283**	**768 766**	**767 038**	**760 470**
Fire and Elements of Nature	**40 590**	**34 463**	**29 100**	**25 913**	**23 680**	**26 093**	**23 120**
Other damage in property	**665 283**	**697 713**	**715 188**	**739 370**	**745 086**	**740 945**	**737 350**
Agricultural	**21 478**	**24 244**	**24 240**	**24 143**	**22 874**	**19 485**	**17 982**
Agricultural - Fire	447	402	562	617	313	261	293
Agricultural - Crops	21 031	23 842	23 678	23 526	22 561	19 224	17 689
Cattle	**139**	**101**	**70**	**46**	**63**	**49**	**61**
Theft	**6 051**	**6 509**	**6 334**	**5 776**	**5 611**	**5 510**	**4 837**
Cristals	**514**	**462**	**415**	**400**	**373**	**299**	**251**
Deterioration of Refrigerated Goods	**131**	**124**	**154**	**80**	**38**	**47**	**36**
Machine Malfunction	**17 774**	**20 669**	**19 508**	**18 311**	**19 648**	**18 435**	**17 495**
Multiple Risks	**577 466**	**594 445**	**615 643**	**637 928**	**652 923**	**663 935**	**670 172**
Multiple Risks - Habitational	356 757	372 529	388 165	404 934	418 066	431 664	438 163
Multiple Risks - Commerce	137 744	130 516	133 316	131 503	144 556	137 865	133 532
Multiple Risks - Industrial	74 731	81 348	83 054	89 482	77 629	81 645	86 663
Multiple Risks - Others	8 234	10 053	11 107	12 008	12 672	12 761	11 814
Others	**41 729**	**51 159**	**48 823**	**52 686**	**43 556**	**33 186**	**26 516**

U: Thousands of Euros

Table 1: Annual Written Premium distribution in Portugal for FODP LoB. Data collected from the report "Produção Anual de Seguro Direto 2013" compiled by Associação Portuguesa de Seguradores (APS)

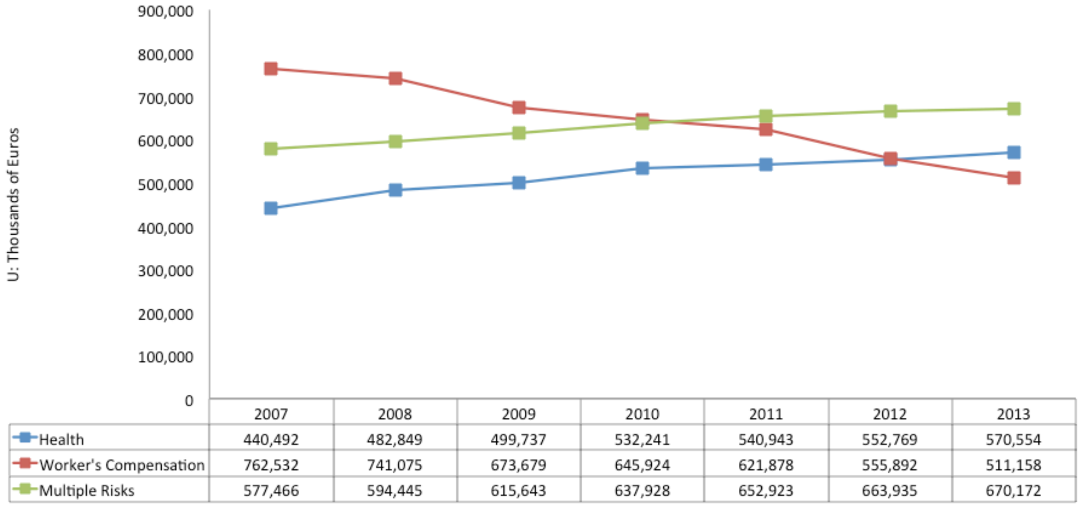
This table shows that the Multiple Risks (MR) branch is the main force behind the FODP LoB, with the 2013 premium collection arising to 670,172 thousands of euros, about 88.1% (670,172 /760,470) of the LoBs written premiums.

For comparison sake, if we also split the AH LoB into its branches and analyse their respective annual premiums and weights on the LoB, we get the results shown in the table below:

	2007	2008	2009	2010	2011	2012	2013
Accidents&Health	**1 372 230**	**1 396 233**	**1 353 441**	**1 356 545**	**1 321 837**	**1 262 348**	**1 232 461**
Accidents	**931 739**	**913 384**	**853 703**	**824 304**	**780 894**	**709 580**	**661 907**
Workers compensation	762 532	741 075	673 679	645 924	621 878	555 892	511 158
Personal Accidents	156 747	161 895	171 530	174 912	156 219	151 588	149 098
Transportated people	12 459	10 414	8 495	3 468	2 797	2 100	1 650
Health	**440 492**	**482 849**	**499 737**	**532 241**	**540 943**	**552 769**	**570 554**

Table 2: Annual Written Premium distribution in Portugal for AH LoB. Data collected from the report "Produção Anual de Seguro Direto 2013" compiled by Associação Portuguesa de Seguradores (APS)

In 2013, the Workers compensation (WC) sub-branch was responsible for 41.5% of the total LoB premiums and the Health branch for 46.3%. Together, these two branches are responsible for 87.8% of the AH LoB premiums, similar to the weight of the MR branch in the FODP LoB (88.1%). Now we can analyse these three branches premium behaviour in the last 7 years and get a different picture from the one in Figure 1:

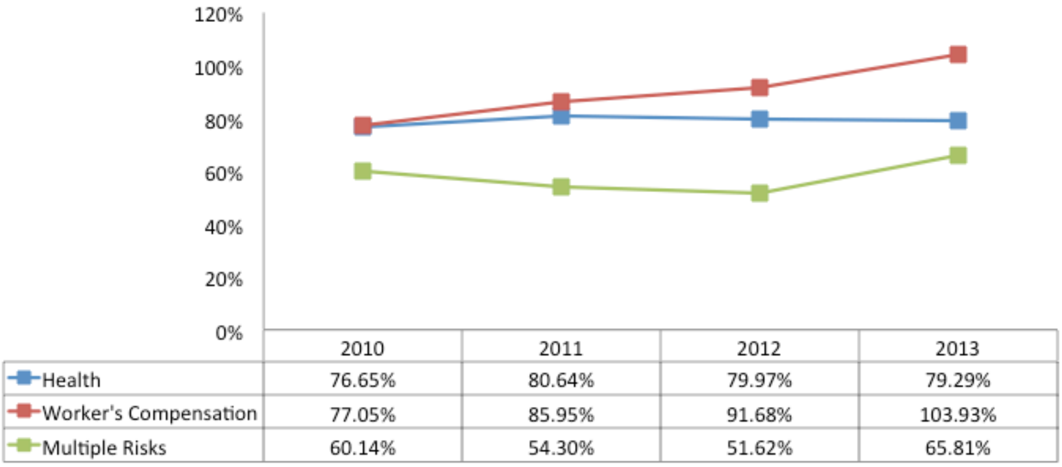


	2007	2008	2009	2010	2011	2012	2013
Health	440,492	482,849	499,737	532,241	540,943	552,769	570,554
Worker's Compensation	762,532	741,075	673,679	645,924	621,878	555,892	511,158
Multiple Risks	577,466	594,445	615,643	637,928	652,923	663,935	670,172

Figure 2: Premium collection in Portugal for Worker's compensation, Health and Multiple Risks. Data collected from the report "Produção Anual de Seguro Direto 2013" compiled by Associação Portuguesa de Seguradores (APS)

The period under analysis emcompasses the arise of the Portuguese economic crisis which lead to some losses in the Non-Life business. The decline in premium collection is clear for a number of LoBs, especially those that are more sensitive to the macroeconomic conjuncture. While the increase in the Portuguese unemployment rate helps to explain the WC premium decline, for MI two axis must be analysed: in the mandatory third party liability cover, the decline is partially justified by a fierce market competition, especially from the direct companies that manage to lower their insurance premiums by using direct communication means with their costumers, mainly the internet, saving in paperwork and labor force. In the optional covers, the economic conjuncture again plays a role, with families trying to save money wherever they can.

However, these social and economical factors don't seem to have such an impact in the MR and Health branches. The following figure showing the behaviour of the loss ratio for these branches, helps to prove our point:



	2010	2011	2012	2013
■ Health	76.65%	80.64%	79.97%	79.29%
■ Worker's Compensation	77.05%	85.95%	91.68%	103.93%
■ Multiple Risks	60.14%	54.30%	51.62%	65.81%

Figure 3: Loss ratio for the Health and WC branches and the MR sub-branch. Data collected from the reports "Produção Anual de Seguro Direto 2013" and "Variáveis Trimestrais 2013.12" compiled by Associação Portuguesa de Seguradores (APS)

The rise of the loss ratio for the MR branch in 2013 cannot be dissassociated from the extreme weather conditions in winter and summer of that year, a tendency that will probably aggravate in the future. The WC premium decrease, as a consequence of a decrease in the employed labor force, lead to losses over 100% of the writtens premiums. The Health businnes is stable, certainly the most crisis-resistant among these three, showing that the Portuguese people have increased their confidence in private health insurances, wether resulting of a free choice or of the attaint of the public health system.

In the second half of the 1980's, the MR branch started to make its way into the Portuguese insurance market. It started as a more complete alternative to the traditional fire insurance, since insurance companies added to the fire cover a set of optional covers that also reflected potential risks in losses of property. The insurance was sold as a package of covers, wich made it cheaper and easier for the consumer, who didn't have to underwrite several distinct insurance policies. Each cover in this package has its guarantees and exclusions, wich differ from company to company. The following table shows the most common covers in a MR contract and a very brief description of its guarantees and exclusions:

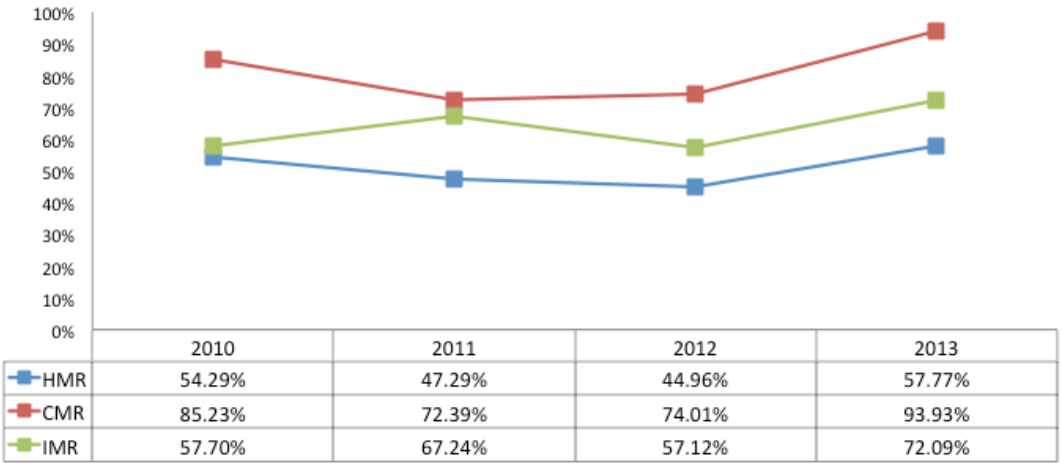
Cover	Guarantees	Exclusions
Vandalism	Damage to insured property directly caused by acts of vandalism, including those resulting from fire and explosion.	robbery and theft of the insured goods.
Land Subsidence	Landslides, mudslides and land subsidence.	Damage resulting from the partial or total colapse of a structure not related with geological risks.
Water Damage	Rupture, malfunction, clogging or overflow of internal network of water supply and sewerage.	Taps left open.
Deterioration of products	The cost of the deterioration of certain products in the commercial activity of the insured.	Coverage is limited to the products expressed in the contracted insurance.
Theft or Robbery	Disappearance, destruction or damage to insured property as a result of theft or robbery.	Burglary and theft in non-permanent housing.
Strikes, Riots, Public Disorder	Damage to insured property directly caused by acts of strikes, riots or public disorder.	robbery and theft of the insured goods.
Fire, Lightning or Explosion	Damage of insured goods as a result of the fire and of the means employed to fight the fire.	The simple action of heat (no flames).
Floods	Torrential rain (> 10 mm /10 min), bursting of dikes and dams.	Direct sea action, infiltrations trough walls.
Others	Other damage to property.	-
Loss of profits or income	The fixed costs that the insured will have to continue to endure despite the complete or partial interruption of activity, as a result of an accident (salaries, insurance premiums, taxes, depreciation and other fixed costs attributable to the exercise).	Coverage is limited to the risks or activities expressed in the contracted insurance.
Broken glass	Break or fracture of glass plates and fixed mirrors, marble stones and other fixed decorative stones, as well as fixed bathroom fixtures .	Damages resulting from defect or manufacturing defect, placement, assembly or disassembly and resulting from inadequate support.
Third party liability	Compensation for pecuniary and non-pecuniary damage the insured is legally required by a third party.	Coverage is limited to the risks or activities expressed in the contracted insurance.
Electrical Risks	Damage to electrical machines, transformers, apparatus and electrical systems and accessories, as a result of direct effects of electric current.	Damage to fuses, heaters, lamps and cathode ray tubes of electronic components.
Seismic Risks	Direct action of earthquakes, volcanic eruptions, tsunamis and fire resulting from these phenomena.	Buildings that have not been scaled according to the regulations in force at the date of construction and the structure.
Storms	Typhoons, cyclones, overflows caused by rainfall, snow or hail.	Direct sea action, infiltrations trough walls.

Table 3: Multiple risks covers and its guarantees and exclusions

Soon, the market behaviour of the product lead to the creation of a "basic kit" of covers wich usually includes Fire, Theft or Robbery, Floods, Storms and Water damage.

The premium/sum insured calculations of a policy with several covers that reflected such distinct social and natural hazards can be done in several ways: Wether the premium/sum insured of a policy is the sum of the several premium/sum insured of the different covers, or a total premium/sum insured is calculated for the whole of the basic kit and the remaining covers have separate calculations for premium/sum insured. Usually the seismic risks cover is reinsured and treated apart from the other covers.

As shown in Table 1, the MR branch is divided into four sub-branches: Habitational Multiple Risks (HMR), Commerce Multiple Risks (CMR), Industrial Multiple Risks (IMR) and Other Multiple Risks (OMR). All of these branches share the same diversity of exposed risks, materialized in the set of covers each policy possesses. Naturally, the most affected covers, in terms of claim severity or claim frequency, vary in each sub-branch. There is a mandatory fire insurance or a MR insurance with the fire cover for every household in horizontal property in Portugal, but in commerce and industry that obligation is not present and the entrepreneur chooses freely wich covers he desires to protect his business. This sometimes leads to the celebration of "tailor-made" contracts where the company directly negociates the insurance contract with the client. This helps to explain the more dispersed loss ratio on IMR and CMR as opposed to the more stable one of HMR.



	2010	2011	2012	2013
HMR	54.29%	47.29%	44.96%	57.77%
CMR	85.23%	72.39%	74.01%	93.93%
IMR	57.70%	67.24%	57.12%	72.09%

Figure 4: Loss ratio for the HMR, CMR and IMR sub-branches in Portugal. Data collected from the reports "Produção Anual de Seguro Direto 2013" and "Variáveis Trimestrais 2013.12" compiled by Associação Portuguesa de Seguradores (APS)

We now arrive to the object of our practical experiment in this work, wich is the IMR sub-branch. Policies in this sub-branch cover a wide variety of economical activities and therefore are, to some extent, the reflex of the Portuguese Industrial tissue. The typical Portuguese industry is a small family-based business, still, a few large industrial groups dominate the market. This economical assimetry is reflected in the behaviour of the policies covering these risks, with the great majority of the policies generating small claim sizes and only a handful of policies generating the greatest part of the sub-branches total loss.

The Portuguese industries with greater production value include the food, drinks and tobacco industry; the water/gas production/distribution industry and the textile industry[2]. This sector lost economical relevance, specially since the country joined the European Union and applied more efforts on the growth of the tertiary sector, leaving the primary and secondary sectors behind. This lead to a serious

[2] http://www.centromarca.pt/folder/conteudo/620_Indústria%20Portuguesa_CIP_Relatório%20Final_AMA.pdf

lack of competitivity in the global market, but still, in 2011, the Industrial sector contributed with 24%[3] of the Gross Domestic Product and in 2010, also with 24% of the country's employed labour force. The recovery of this economical sector is vital for the economical recovery in itself and it must certainly take into account a more environmental friendly approach and a greater technological knowledge.

3. Description of the data

The data used in the next chapters was collected by APS from 7 different insurance companies operating in Portugal in 2010 and 2011 that exploit the IMR sub-branch, with reference date 31.05.2012. These 7 companies possessed all together a market-share of 70.7% of the IMR market. Policies in force at least one day in 2010 were used to build the models described in chapters 4 and 5. These models were then used to predict claim values for the policies in force at least one day in 2011, and these prediction were compared with the observed claim amounts in 2011. The following table summarizes some descriptive statistics of the samples.

Year	Number of policies exposed	Annualised exposure	Number of claims	Total Loss
2008	14,810	14,648	2,501	23,093,043€
2009	15,732	15,033	3,130	31,139,744€
2010	18,505	15,756	3,167	31,358,004€
2011	18,941	15,674	3,036	36,894,403€

Table 4: Descriptive statistics of the experiment samples

For each policy in the study, the following possible explanatory variables were recorded:

[3] Taken from "Principais desafios da indústria em Portugal – 2013. Uma abordagem coerente para a dinamização do sector". http://www.pwc.pt/pt/publicacoes/imagens/2013/pwc_principais_desafios_industria.pdf.

Variable	Type of variable	Codification in model	Type of factor
Covers composing the policy	Binary/Dummy	C_i, i=1, …, 15	Covers factor
Deductibles composing the policy	Binary/Dummy	F_i, i=1, …, 15	Deductible factor
Economical Activity Code (CAE 3.0)	Categorical	CLASS_CAE (4 classes)	LoB factor
Sum insured Class	Categorical	SUM_INSURED_CLASS (7 classes)	Sum Insured factor
NUTSIII[4]	Categorical	NUTS3 (31 classes)	Regional factor
Exposure years	Continuous	EXPOSURE_YEARS	Exposure factor
Covers composing the policy in the previous year	Binary/Dummy	C_{ij}, i=1, …, 15; j=1	Past covers and deductibles factor
Covers composing the policy two years ago	Binary/Dummy	C_{ij}, i=1, …, 15; j=2	Past covers and deductibles factor
Deductibles composing the policy in the previous year	Binary/Dummy	F_{ij}, i=1, …, 15; j=1	Past covers and deductibles factor
Deductibles composing the policy two years ago	Binary/Dummy	F_{ij}, i=1, …, 15; j=2	Past covers and deductibles factor
Number of claims in the previous year	Categorical	TOTAL_CLAIMS_1_CAT (5 classes)	Past claim behaviour factor
Claim amounts in the previous year	Continuous	TOTAL_LOSS_1	Past claim behaviour factor
Claim Severity in the previous year	Continuous	SEVERITY_1	Past claim behaviour factor
Number of claims two years ago	Categorical	TOTAL_CLAIMS_2_CAT (5 classes)	Past claim behaviour factor
Claim amounts two years ago	Continuous	TOTAL_LOSS _2	Past claim behaviour factor
Claim Severity two years ago	Continuous	SEVERITY _2	Past claim behaviour factor

Table 5: Explanatory variables for the models

[4] http://pt.wikipedia.org/wiki/Unidades_Territoriais_Estat%C3%ADsticas_de_Portugal

The study of the covers and the existence of the respective deductibles is of primary importance. The objective in studying the impact of these variables in claim frequency and severity is to assess the effect of the "tailor-made" contracts referred in the penultimate paragraph of chapter 2. Will the direct negotiation of the covers in the contract withdraw or emphasize the importance of the deductibles? If so, in which covers will that impact be significant? All the covers shown in Table 5 were used as dummy variables, with value 1 if the cover was present in the policy and zero otherwise, except the seismic risks cover. This cover deserves a special treatment, given it's low frequency and high cost, and it's usually reinsured. The order of the covers/deductibles in table 3 is the order of the covers/deductibles in the respective outputs, i.e., C_1 and F_1 refer to the Vandalism cover and deductible; C_{15} and F_{15} to the Storms cover and deductible. There is no C_{14} and F_{14} since that refers to the seismic risks cover.

Also, an understanding of the claim amount and frequency behaviour trough different CAEs is of interest, but here we payed the price of collecting data from 7 different insurance companies. Different companies use different codes in an immense panoply of economical activities, and when they all come together it's not easy to find an algorithm that standartizes them all, since the variable is qualitative. An effort was made to allocate each activity to it's correspondent CAE 3.0[5], but we only achieved satisfactory results with a larger aggregation, shown in the next table:

[5] http://www.ine.pt/ine_novidades/semin/cae/CAE_REV_3.pdf

CLASS	CAE 3.0
1	Administrative activities; Artistic and sporting activities; Communication activities; Health activities; Accommodation and eatery activities; Electricity and gas industries.
2	Transport and storage activities; Extractive industries; Educational activities; Construction activities; Agriculture and animal production activities
3	Unknown
4	Textile industries; Manufacturing industries; Water and waste management; Auto Repair

Table 6: Categorization of the CAE 3.0 variable

Other studied factors include the exposure factor, measured by the proportion of the year the policy was in force; a regional factor, measured by the geographical localization of the risk and categorised in NUTSIII. The Sum Insured factor, measured by the sum insured variable, will help us understand if the size of the industry matters: we expect to see a higher claim frequency and severity in a policy with higher sum insured. This variable was categorized in the following way:

Class	Left limit (open)	Right limit (closed)
1	0	50.000€
2	50.000€	100.000€
3	100.000€	300.000€
4	300.000€	1.500.000€
5	1.500.000€	3.000.000€
6	3.000.000€	20.000.000€
7	20.000.000€	+∞

Table 7: Categorization of the sum insured variable

The past behaviour factor of the policy will also help us understand it’s present behaviour in terms of frequency and severity. This variable was categorized in the following way:

Class	Rule
0	The policy was present in lag 1/2 for a period of more than 180 days with no claims.
1	The policy had 1 claim in lag 1/2
2	The policy had 2 claims in lag 1/2
3+	The policy had 3 or more claims in lag 1/2
absent	In the lag 1/2, the policy was not present, or, it was present for a period less than 180 days with no claims.

Table 8: Categorization of the past claim frequency variable

The same policy can be in force in different places, so the pair (policy,postal code) was considered as the risk to be studied. For each distinct combination of (policy,postal code), a record was created in the data base. That record would then be completed with the values of the other explanatory variables in table 5. An example of the database disposal is given in table 9.

i	Policy_ id	Postal Code	Exposure years	C1	…	C15	F1	…	F15	CAE (3.0)	Sum Insured	…	Number of claims	Claim Size	Claim Severity
1	9999	2855	0.75	1		0	1		0	X	50000		0	0	0
2	9999	2800	0.75	1		1	0		0	X	50000		2	1000	500

Table 9: Database disposal

Using this type of database disposal, for each line we will try to model the response variable “Number of claims” using a GLM with a quasi-poisson density and using the exposure factor as an offset (view section 4.2). We will also model the “Claim severity” using a GLM with a Gamma density, with weights defined by the “Total number of claims”. This is the usual approach or the frequency/severity approach.

The total loss approach consists in modelling the variable “Claim Size” using a GLM with a density from the Tweedie family of distributions.

4. An overview on GLMs

Over the last years GLMs became a common statistical tool to model actuarial data, mainly because the regression is extended to distributions from the exponential family and secondly because a GLM models the additive effects of the explanatory variables on a transformation of the mean , instead of the mean itself. We will present a GLM formulation by McCullagh & Nelder (1989), where a GLM is described by the following assumptions:

- There is a response y observed independently at fixed values of explanatory variables x_1, …, x_p.

In the scope of this work, depending on the context, the response variable represents the number of claims generated by a policy, the claim severity of a policy or the claim size of a policy. The explanatory variables represent the variables described in table 5. This first assumption aims to isolate the ocurrence of a claim (or it’s severity/size) as independent from the ocurrence (or severity/size) of other claims, thus excluding chain reactions. However, in the reality of the IMR sub-branch and in most of the other LoBs, this might not occur. A storm might boost broken glass or water damage. A seism might trigger other natural or social hazards such as fire or vandalism. The way companies themselves deal with a claim might also subvert reality, if for instance, two dependent claims of storm and broken glass are reported as one unique claim or vice-versa. An effort to allocate each claim to its affected cover(s) might seem pointless work at the time of the expertise, but it will make a great difference for the actuary who will work with the data later on.

- The distribution of y has density of the form:

$$f(y_i; \theta_i; \phi) = exp\left[\frac{(y_i\theta_i - b(\theta_i))}{a_i(\phi)} + c(y_i, \phi)\right] \qquad (1)$$

for a positive parameter ϕ and suitable functions $a_i(.)$, $b(.)$ and $c(.,.)$. Usually $a_i(\emptyset)$=$\frac{\emptyset}{w_i}$, where w_i is a set of know weights (see section 4.2) and the domain of each θ is an open interval satisfying b(θ)<∞. Some well know results using this framework are: E$(y_i) = \mu_i = b'(\theta_i)$ and Var(y_i) = ϕV(μ_i) where V(.) = $b''(\theta_i)$ is the variance function for each observation.

- The explanatory variables may only influence the distribution of y through a single linear function called the linear predictor $\eta = \beta_1 x_1 + \cdots + \beta_p x_p$, where the β parameters are derived by maximizing the log-likelihood defined as:

$$l(\beta, \phi) = \sum_{i=1}^{n} \ln f(y_i, \beta, \phi) = \sum_{i=1}^{n} \{\ln c(y_i, \phi) + \frac{y_i\theta_i - b(\theta_i)}{a_i(\emptyset)}\} \qquad (2)$$

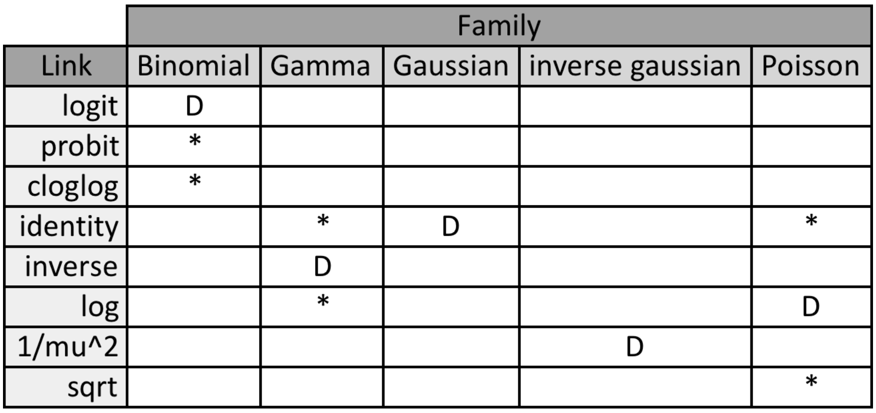
The maximum lilkelihood estimation of the β_i parameters won't be presented here but can be consulted in, e.g, De Jong, P. And Heller, G.Z. (2008).

- The mean μ is a smooth invertible function of the linear predictor:

$$\mu = g(\eta) <=> \eta = g^{-1}(\mu) = \mathrm{h}(\mu) \qquad (3)$$

where the function h(μ) is called the link function.

The link function will depend on the choice of the family for the response variable y and its choice is suggested by the functional form of the relationship between y and the explanatory variables. In the table below we present the most common link functions for some important families of distributions:



	Family				
Link	Binomial	Gamma	Gaussian	inverse gaussian	Poisson
logit	D				
probit	*				
cloglog	*				
identity		*	D		*
inverse		D			
log		*			D
1/mu^2				D	
sqrt					*

Table 10: Families and Link function. The canonical (or default) link is denoted by D, while * denotes other possible links for the family

We now take the example of the Poisson family, which is theoretical foundation for the claim frequency model, even if some corrections will have to be made, as we will see in sections 4.1 and 4.2.

For a Poisson distribution with mean μ, applying the default link function log to the Poisson density f(.), equation (1) becomes:

$$\log\big(f(y)\big) = ylog(\mu) - \mu - \log\,(y!) \tag{4}$$

so $\theta = log(\mu)$, $\phi = 1$, and $b(\theta) = \mu = e^{\theta}$. This makes sense with the well known result $\mathrm{E}(y_i) = Var(y_i)$ for the Poisson model since, as we've seen above, Var(y_i)= ϕ×V(μ_i)= 1×V(μ_i). However, in many empirical analysis, data appears more dispersed than expected ($\phi > 1$). This phenomenom is called over-dispersion and we adress it in the next section.

4.1 Over-Dispersion in Poisson GLMs

Over-dispersion (under-dispersion) results when the data appear more (less) dispersed than expected under the Poisson model. Under-dispersion is a rare and not so interesting case for our work and won't be treated here. Venables W.N and Ripley B.D (2002) are among the many who have allready adressed this problem, wich can be tackled it in different ways:

One consists in introducing some variability in the Poisson mean λ by assuming it follows a certain distribution. This mixture of distributions can be done in different ways with different solutions, the most usefull being the case when λ is gamma-distributed. This mixture provides us with the negative-binomial regression model, a parametric way of modelling over-dispersion. Apart from section 7.4 of Venables W.N and Ripley B.D (2002) we highlight a paper by Ismail and Jemain (2007) "Handling over-dispersion with Negative Binomial and Generalized Poisson Regression Models".

Another way to approach the over-dispersion problem is to consider $\emptyset$ as a parameter to be estimated and to use quasi-likelihood – see among others McCullagh and Nelder (1989) and De Jong, P. And Heller, G.Z. (2008) - instead of likelihood, since we are not using a distribution to estimate the model (ϕ is no longer a constant equal to 1); we are estimating the model based on the definition of the first two moments of Y_i.

We can detect over-dispersion if the magnitude of the residual deviance is much greater than the residual degrees of freedom in the fitted model. Another way of doing it is to fit a GLM using the quasipoisson family and compare the estimate of the dispersion parameter $\emptyset$ with theoretical value 1.

4.2 Offsets

As presented by De Jong, P. And Heller, G.Z. (2008), offsets are used to correct for group size or, as it is the case in this work, to correct different time periods of observation. Some policies in our database were exposed trough the whole year of 2010, some only six months, others only one day, and naturally this point has to be taken into account. The exposure w ("Exposure years" in table 4) was measured as a proportion of the year, thus with maximum value 1. We will assume time homogeneity, i.e, we will model the ocurrence rate μ/w, where μ is the mean of the count y. From the second assumption in chapter 4, $\ell(\mu/w)= \beta_1 x_1 + \cdots + \beta_p x_p$, and when ℓ is the log function, we get $\ln(\mu/w)= \beta_1 x_1 + \cdots + \beta_p x_p \Rightarrow \ln(\mu)=\ln(w)+ \beta_1 x_1 + \cdots + \beta_p x_p$ where $\ln(w)$ is called an "offset". An offset is effectively another x variable in the regression, with a given β coefficient equal to one. Using the offset, y has expected value directly proportional to exposure: $\mu = we^{x'\beta}$.

However, our assumption didn't show adherence to the data. Setting the annualised exposure as an offset provided a statistically significant negative parameter for the regressor EXPOSURE_YEARS, forcing us to put aside the time homogeneity hypothesis. As an alternative, we used the variable EXPOSURE_YEARS as a regressor in the model, but no offset.

5. The Tweedie subclass of distributions

The actuarial modelling universe gained a new powerfull tool when Maurice Twedie published "An index that distinguishes between some important exponential families" in 1984. This paper presented a new subclass of exponential dispersion families, suitable to use in GLMs. This approach has gained great popularity among actuaries, with other interesting papers by Smith and Jørgensen (2002) or Kaas (2005), where this distributions achieved very satisfactory results modelling insurance premiums in a GLM framework.

The biggest problem in modelling total claim amounts with data from individual policies, is that most of the losses generated are zero, and for the policies with a positive loss the data is highly skewed. The typical way to overcome this problem consists in working with separate models for frequency and severity, but since the Tweedie distribution can be parametrized as a Compound Poisson distribution (Smith and Jørgensen (2002)), with a probability mass at zero, the whole data can be modeled at once, using the total loss of a policy as the response variable in the GLM. Of course both approaches make use of very debatable assumptions in terms of frequency/severity correlation.

An exponential dispersion family (defined in equation (1)) is a Tweedie Family if the domain of its variance function V is $[0,\infty[$ with

$$V(\mu) = \mu^{p}, \text{ for some } p \in \mathbb{R}. \tag{5}$$

The Tweedie families encompass some well known distributions that are characterized by the value of the parameter p. The following table presents some well know distributions that can be seen as the Tweedie family for different values of p.

Value of p	Distribution
p=0	Normal
p=1	Poisson
p∈[1,2]	Compound Poisson-Gamma
p=2	Gamma
p=3	Inverse Gaussian

Table 11: Distributions as a function of the Tweedie p parameter

For the remaining values of p, the Tweedie families characterize distributions that are supported on ℝ. For p>2 it characterizes distributions that have support in [0,∞[, and for p∈]0,1[there is no probability measure. For the purpose of this work we will focus on the case p∈[1,2] wich characterizes a Poisson-Gamma distribution, using a log link function in order to work with a mulitplicative model. This link is also usefull in the sense that the parameter signals will be equivalent to the effect signal, with a positive parameter showing greater risk.

The Tweedie distribution therefore accommodates the parameter λ from the claim count distribution and the parameters θ and α from the claim size distribution into its own parameters μ, ϕ and p. Smith and Jørgensen (2002) translate the parameters of the Compound Poisson Distribution into the usual Tweedie parameters in the following way:

$$\mu = \lambda \times \alpha \times \theta \tag{6}$$

$$\phi = \frac{\lambda^{1-p} \times (\alpha\theta)^{2-p}}{2-p} \tag{7}$$

$$p = \frac{\alpha+2}{\alpha+1} \tag{8}$$

From equation (6) we can see that the expected value of a Tweedie distribution takes into account the effects of the Poisson and Gamma distributions. In equation (8) we can see that p - the parameter that will define the variance function - will be between 1 and 2 and depends only on the shape parameter of the claim severity distribution. The dispersion parameter calculated in equation (7) will take into account the effects of the Poisson and Gamma distributions and parameter p.

6 Results

All the models were fitted using R. When possible, variables in the models were selected using the backward selection procedure and their names are coherent with table 5. For each policy in the study, the explanatory variables represent:

- A past claim behaviour factor (8 parameters)
- A Sum Insured factor (6 parameters)
- A covers factor (14 parameters)
- A deductibles factor (14 parameters)
- A regional factor (30 parameters)
- An exposure factor (1 parameter)

- A LoB factor (4 parameters)

Statistitians know that a model is only as good as the data it is fitting. We must not forget that when we're modelling the behaviour of a certain variable, we're modelling that behaviour in respect to the collected data and not to reality. To do that, we would have to collect all the availabe data where that variable is in play without errors. Since all the parameter estimates our model gives us are calculated from the collected data, they're not the effective parameters observed in nature. In fact, we will never truly know those unachievable, almost "esoteric" parameters, unless we accurately measure everything everywhere. All we have is an approximation, so, caution analysing model estimates is always in order.

Sections 6.1, 6.2 and 6.4 refer to the modelling of the whole sample. In section 6.3 we used a treshold in the sum insured variable to split the sample into more homogeneous groups in terms of claim frequency and severity. We then experimented the same diferent approaches used to model the whole sample in both groups. In the sections below we will analyse the individual behaviour of the significant variables in each of the models as well as the goodness of fit of each model. In chapter 7 we will compare the predictions all these aproaches produced with the observed 2011 claim values.

6.1 The frequency model

The output of the selected model is shown in annex 1, where TOTAL_CLAIMS – defined as the total number of claims reported in 2010 for each risk - is the response variable. The model captured in an expected way the effects of the Sum Insured factor and the past claim behaviour factor. The regional factor effect was also captured in an expected way, discriminating the regions with a known industrial force or exposed to more severe weather conditions. The LoB factor didn't play an important part in the modelling for the reasons allready adressed in chapter 3. The exposure factor behaviour was somehow surprising, since we were expecting time homogeneity and that wasn't the case. Below, we individually analyse the behaviour of these factors:

- **Past claim behaviour factor**

It is common knowledge to believe that between a policy that has never produced a claim and a policy that produces recurrent claims over the years, the latter is more likely to produce claims in the future. The total claim number per policy was recorded for the previous year (lag 1) and for two years ago (lag 2), and was categorized as shown in table 8.

Prior to the model fitting, the "common knowledge" belief stated in the beggining of this section was supported by the results in following table, taken from the studied database:

		Number of claims in 2010			
		0	1	2	3+
Number of claims in 2009	absent	90%	7%	2%	1%
	0	92%	6%	2%	1%
	1	77%	17%	5%	2%
	2	68%	16%	9%	8%
	3+	52%	13%	14%	21%

Table 12: Double entry table for the study of past claim frequency (lag 1)

It is clear that in a 1-year lag, past claim frequency has an influence in present claim frequency. If we make the same exercise for lag 2 we can see the past policy behaviour effect is still present, even if slightly mitigated:

		Number of claims in 2010			
		0	1	2	3+
Number of claims in 2008	absent	89%	7%	2%	1%
	0	92%	6%	2%	1%
	1	76%	15%	7%	2%
	2	68%	15%	8%	8%
	3+	48%	12%	20%	20%

Table 13: Double entry table for the study of past claim frequency (lag 2)

The results of a 95% confidence interval for lag 1 are shown in the following tables (these, and the remaining parameter values/model information can be found in the respective annexes):

Class	Left limit	**Parameter Value**	Right Limit
absent	-0.259807856	**-0.12199**	0.01583404
1	0.320172390	**0.48466**	0.64913786
2	0.772898457	**0.97291**	1.17292272
3+	1.467012439	**1.66566**	1.86431642

Table 14: Parameter values for the past claim frequency explanatory variable in lag 1

This is an usefull view, in the sense one can immediately assess if the parameter value is different from zero and if the classes are sufficiently far apart. With class 0 as the reference class, we conclude that a policy in this class has the same risk profile as an absent policy since zero is included in the respective confidence interval. Troughout the classes, the increase in the parameter estimates makes sense with the values in table 12, showing that an increase in past claim frequency has a positive effect in present claim frequency.

Let's make the same exercise for lag 2:

Class	Left limit	**Parameter Value**	Right Limit
absent	0.130063480	**0.25276**	0.37544970
1	0.324137406	**0.51873**	0.71332255
2	0.550084345	**0.83310**	1.11611561
3+	0.425747502	**0.74155**	1.05735813

Table 15: Parameter values for the past claim frequency explanatory variable in lag 2

As expected, the importance of past claim frequency attenuates with higher lag values. With class 0 again as the reference class we now observe some overlapping in the confidence intervals. The 3+ class is entirely included in class 2, wich overlaps with class 1. Given these results, a better way to assess the effect of the 2-year lag claims in present claim frequency would be by relaxing the categorization of the variable, using a binomial categorization with value 0 if there were no claims and value 1 if there were claims.

- **Sum Insured factor**

The categorization of the sum insured is defined in table 7. We're expecting lower claim frequency in smaller industries and higher claim frequency in larger industries, hence, if the sum insured in each policy is adequately defined, we're expecting higher claim frequency in higher sum insured values. The following table shows the estimates and p-values for the observations in these classes for model 1.

Class	Left limit	**Parameter**	Right limit
2	-0.038417628	**0.29239**	0.62319222
3	0.246971216	**0.52405**	0.80112050
4	0.805494042	**1.06629**	1.32708815
5	0.815743440	**1.08898**	1.36221172
6	1.385471297	**1.64695**	1.90843234
7	1.920151475	**2.21847**	2.51679793

Table 16: Parameter values for the sum insured explanatory variable.

Even if some overlapping is observed in the confidence intervals, the claim frequency increases with the sum insured as expected.

- **Covers factor**

The next table shows us the covers that proved to be statistically significant in the claim frequency model:

Cover	Left limit	**Parameter**	Right limit
Theft or robbery (C_5)	0.788835669	**1.15922**	1.52959632
Floods (C_8)	0.551425612	**0.90353**	1.25563666
Others (C_9)	-0.984871302	**-0.61404**	-0.24321789
Loss of profit or income (C_{10})	0.167899325	**0.31676**	0.46562108
Third Party Liability (C_{12})	-0.545444410	**-0.33003**	-0.11461460
Electrical risks (C_{13})	0.529752493	**0.65102**	0.77228688

Table 17: Parameter values for the significant covers.

At a 95% confidence level, signing the covers 5, 8, 10 and/or 13 will increase the risk of a policy. Signing covers 9 and 12 will probably result on the opposite effect.

- **Deductibles factor**

The next table shows us the deductibles that proved statistically significant:

Deductible	Left limit	**Parameter**	Right limit
Deterioration of products (F_4)	-0.838085286	**-0.51006**	-0.18202561
Theft or robbery (F_5)	-1.089454102	**-0.82172**	-0.55398206
Floods (F_8)	-0.829839379	**-0.61881**	-0.40778411
Others (F_9)	0.275408821	**0.45520**	0.63498633
Loss of profit or income (F_{10})	0.518603751	**0.77362**	1.02864231
Third Party Liability (F_{12})	0.270621994	**0.48715**	0.70366821

Table 18: Parameter values for the significant deductibles in model 1.

The existence of deductibles for covers 4, 5 and 8 originates an expected negative signal, since we expect deductibles to lower claim frequency. However, the existence of deductibles for covers 9, 10 and 12 increases claim frequency. In an actuarial/economical point of view this doesn't make much sense but it might be explained by a series of factors: perhaps the "tailor-made" contracts referred in chapter 2, where the policyholder can directly negotiate the signed covers and respective deductible values with the insurance company withdraws importance from the deductibles. Or, maybe for these covers, a higher value of the deductible is negotiated in comparison with other covers.

- **Regional factor**

NUTS3	Left limit	**Parameter**	Right limit
Alto Trás-Os-Montes	0.007318649	**0.67591**	1.34450529
Baixo Vouga	0.087309393	**0.66727**	1.24722914
Cávado	0.015282944	**0.59772**	1.18014867
Dão-Lafões	0.115851513	**0.72057**	1.32528454
Região Autónoma da Madeira	0.232878189	**0.92375**	1.61461770
Serra da estrela	0.522845607	**1.31494**	2.10704034

Table 19: Parameter values for the significant regions in model 1.

In the northern areas of Baixo Vouga, Cávado, Dão-Lafões and Alto Trás-Os-Montes there is a greater concentration of food/drinks industries and manufacturing industries (pottery, paving, sanitaryware, kitchenware and furniture), so it's no surprise to see these 4 neighboring areas representing the same higher risk profile (view footnote 1 in chapter 2 for more information on this subject). The higher risk profile for the Região Autónoma da Madeira zone came as a surprise. This is a zone with little industrial production, even tough it's more specialized in agricultural/food and extractive indutries. The 2010 Madeira storms in february probabily had an influence in the size of this parameter. Finally, Serra da Estrela shows the higher risk profile in claim frequency. In the highest point of continental Portugal the textile/leather and food/animal industries are the most represented, and it also comes as no surprise for this region - so over-exposed to extreme weather conditions - to show the greatest claim frequency among all others.

- **Exposure factor**

Left limit	**Parameter**	Right limit
0.556749585	**0.74814**	0.93953223

Table 20: Parameter values for the significant regions in model 1.

The parameter in table 20 shows us that the claim frequency isn't proportional to time t, but to $t^{0.74814}$ (view section 4.2).

- **LoB factor**

This factor did not prove significant. The reasons for this were allready adressed in chapter 3, after table 5.

- **Overall goodness of fit**

Since we are using a quasipoisson family, there is no likelihood and so the likelihood ratio test is impossible to perform. However, we can perform the chi-square test in terms of the deviances. Using this test, a model is innefective for use if the test statistic (Null Deviance - Residual Deviance) is smaller than the value of the chi-square distribution with degrees of freedom equal to the difference in degrees of freedom between the Null model and the chosen model. In this model we observed:

(16,426-12,349=4,077) $> \chi^2{}_{(19068-19008=60);95\%} \approx 43.19$

And so the model contributes to explain claim frequency better than an overall mean.

In recent years, some data mining techniques have gained popularity in assessing a model's overall goodness of fit. The lift chart is one of them and provides a more visual and intuitive alternative. It is computed in the following way:

a) For each observation take the fitted values, the value of the response variable and the exposure.

b) Order the observations increasingly with respect to the fitted values

c) Divide the ordered data in groups that have equal number of observations

d) Plot the mean of the response variable, the mean of fitted values an the total exposure for each group

The lift charts for the claim frequency model is shown below:

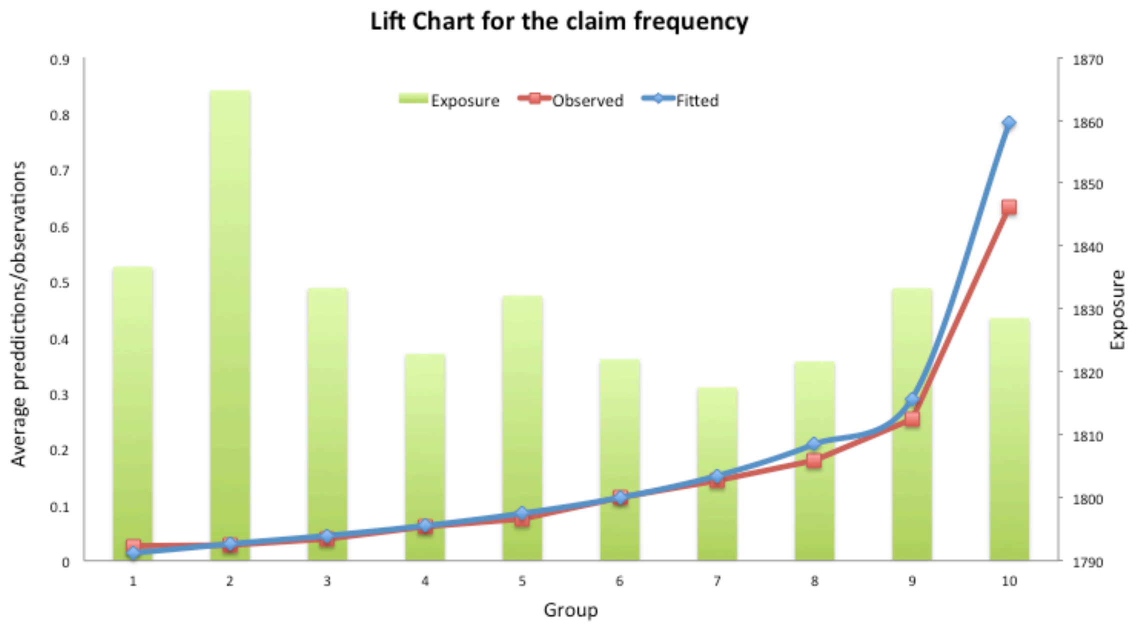


Figure 5: Lift chart for the claim frequency model

This kind of graph gives information about two aspects of the model. On the one hand, by seeing the trend of the curve for the observed means it is possible to see if the model more or less identifies the groups that have greater claim frequency. In the other hand, the vertical distance between the predicted mean and the observed mean gives the idea of how far the predictions are from the observations. Observing the graph, we can expect a bit of over-estimation in the highest risk groups.

6.2 The severity model

In the framework of the frequency/severity models, when it comes to the severity part of the problem, we use the claim severity per policy as an observation from the response variable. The fitted model is weighted with the total number of claims per policy and later is compounded with the fitted model for the total number of claims per policy. The output for this model in shown in annex 2 where CLAIM SEVERITY is the response variable, defined as the average cost of the claims reported in 2010 for each risk. Below, we give an outlook in the statistically significant variables in this model:

- **Past claim behaviour factor**

Class	Left limit	**Parameter Value**	Right Limit
absent	-0.15088905	**0.2004174**	0.551723761
1	-0.17381534	**0.2982403**	0.770295898
2	-0.05733268	**0.5252304**	1.107793390
3+	-1.15651517	**-0.6042307**	-0.051946227

Table 21: Parameter values for the past claim frequency explanatory variable in lag 1

The most interesting fact shown in this table is the negative parameter in the 3+ class. This shows that policies with 3 or more claims in the previous year have lower claim amounts the next year. Classes 1, 2 and absent proved to be not significant in determining future claim amounts, probabily because these policies will generate claims with a small size, since as we can see in annex 1, these policies will generate a greater number of claims in the following year.

- **Sum Insured factor**

Class	Left limit	**Parameter**	Right limit
2	-0.10404119	**0.9222684**	1.948578012
3	0.51066556	**1.3899002**	2.269134810
4	0.59716754	**1.4357131**	2.274258612
5	0.42647161	**1.2917332**	2.156994742
6	0.93338293	**1.7691820**	2.604981060
7	2.20625421	**3.1641212**	4.121988276

Table 22: Parameter values for the sum insured class explanatory variable

Despite a minor irregularity in class 5 (wich can be grouped with class 4), the parameters show a smooth upward trend trough the sum insured classes, until the variable reaches class7 where a big jump is observed. Again, as it happened with the claim frequency model, this class reflects the highest risk profile and a good alternative would probably be to treat it separately from the other classes in order to work with more homogeneous groups and avoid high dispersion in the sample (see section 6.3).

- **Regional factor**

NUTS3	Left limit	**Parameter**	Right limit
Beira Interior Norte*	-0.24376548	**2.2071641**	4.658093682
Lezíria do Tejo	1.45045386	**3.4402621**	5.430070302
Região Autónoma da Madeira	0.19552357	**2.131346**	4.067168460

Table 23: Parameter values for the significant regions in model 1
*** - Significant at 90% level**

Região Autónoma da Madeira again shows a high risk profile in claim severity. As stated in chapter 7.1, it isn't a very industrialized area, but it's exposed to some extreme weather conditions – mainly to storms wich is the most affected cover in terms of claim severity in this zone - that might explain the high parameter value.

Beira Interior Norte is probabily the most unindustrialized region in Portugal, however 15 high claim were enough to assign a 90% confidence high risk profile to this region. Lezíria do Tejo is the portuguese region where claim severity was higher in 2010. This is mainly due to one claim with a cost near 2,000,000€ wich affected the fire cover and is responsible for 63.3% of the total 2010 loss in this region.

- **Covers factor**

Cover	Left limit	**Parameter**	Right limit
Deterioration of products* (C_4)	-0.01198700	**0.7126287**	1.437244321
Broken glass* (C_{11})	-0.97659387	**-0.4584705**	0.059652945

Table 24: Parameter values for the significant covers
*** - Significant at 90% level**

At a 90% level of confidence, signing the Broken Glass cover into a policy isn't expected to generate a higher severity. On the other hand, signing the Deterioration of products cover will probabily do so.

- **Deductibles factor**

Deductible	Left limit	**Parameter**	Right limit
Theft or robbery (F_5)	-0.70108873	**-0.3553347**	-0.009580739
Electrical risks (F_{13})	-1.03463907	**-0.7003232**	-0.366007235

Table 25: Parameter values for the significant deductibles.

Everything as expected in this case, where the only significant deductibles show a negative parameter.

- **LoB factor**

LoB	Left limit	**Parameter**	Right limit
2	-0.62988723	**0.3616341**	1.353155493
3	-0.30632958	**0.7314827**	1.769294996
4	0.14078033	**1.0178882**	1.894996163

Table 26: Parameter values for the significant deductibles.

Class 4 is the only significant class, showing the higher risk profile. This class includes the Textile Industries, Manufacturing Industries, waste/water management and Auto repair.

- **Overall goodness of fit**

Performing the same test as for the frequency model we got the following results that also indicate that the model is a better predicter than the overall mean:

$(10{,}233.8\text{-}6{,}919.7{=}3{,}314.1) > \chi^2_{(1775-1728=47);95\%} \approx 34.76$

We also present the lift chart for the severity model where groups 9 and 10 show a bit of under-estimation.

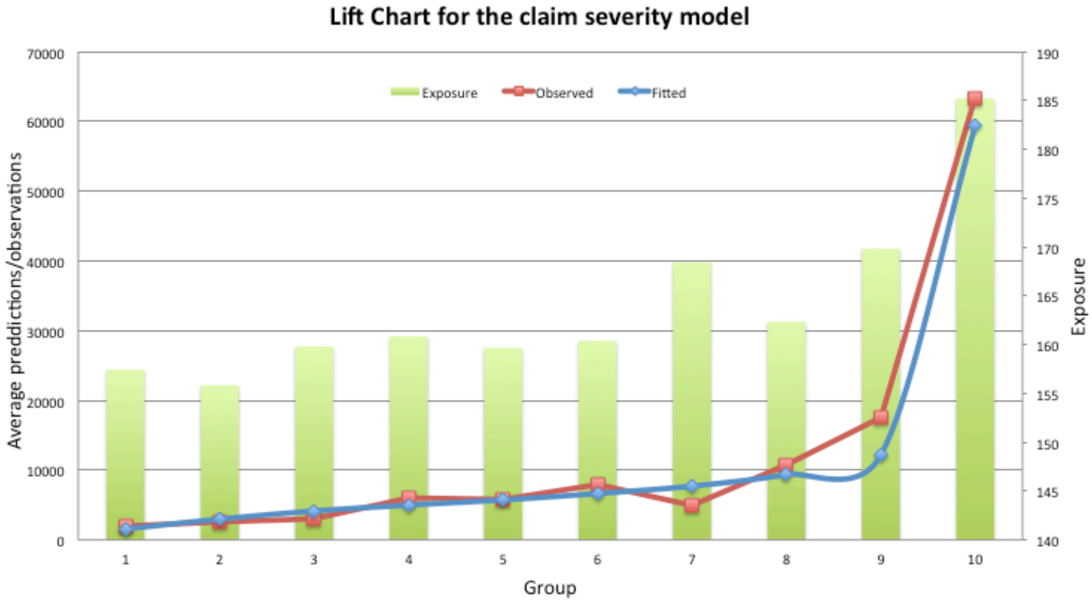


Figure 6: Lift chart for the claim severity model

6.3 Spliting the sample into more homogeneous groups

In the two previous sections we've observed how risk profiles change in respect of the results of a given factor. Stowing all the sample information in only one model implies treating high and low risk profiles under the same rules and such approach may be time-saving but will fail to accurately characterize the different behaviour of those profiles. In this database, the sum insured and the past claim behaviour variables are the most explicit examples of different risk profiles being treated under the same rules. In fact, when we analyse the claim frequency and the average claim cost per sum insured class, we get an obvious picture of how different these diferent risks are:

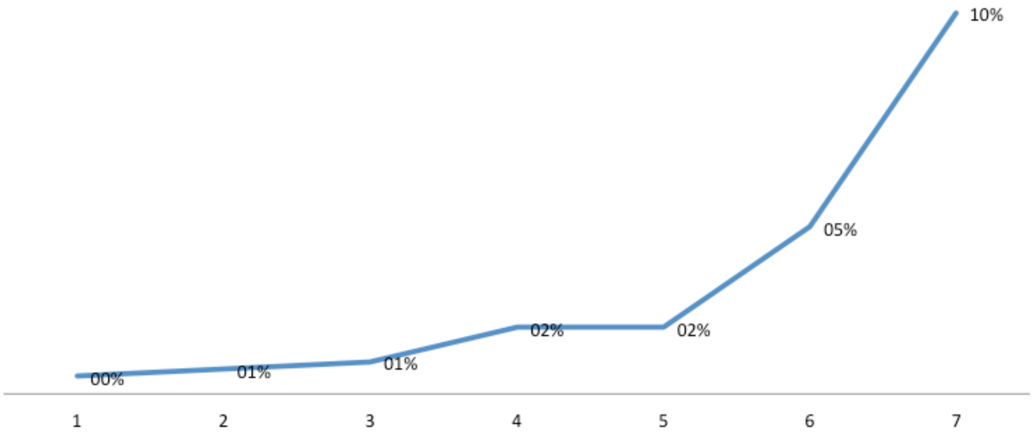


Figure 7: Claim frequency per exposed cover per sum insured class

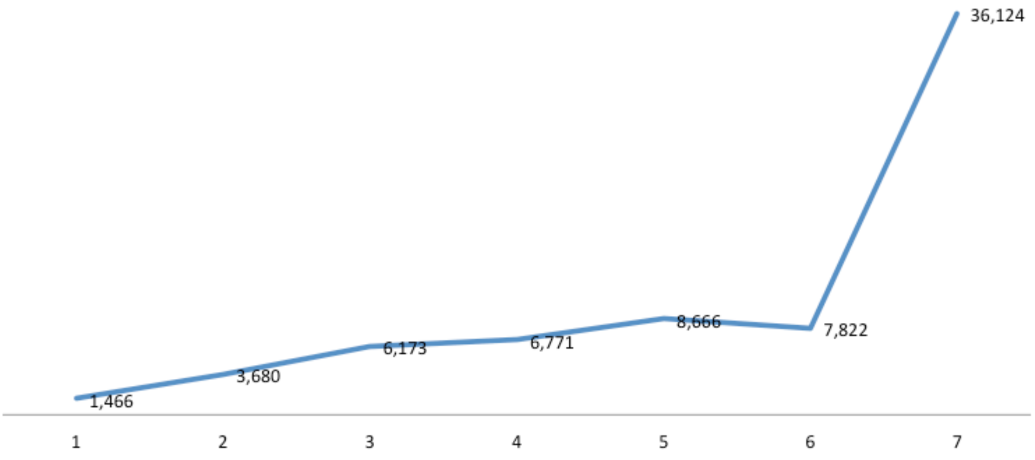


Figure 8: Mean claim severity in € per exposed cover per sum insured class

As shown in table 7, class 7 contains policies with a sum insured greater than 20,000,000€. These are the big Portuguese industries, holding the most expensive materials, machinery and the larger industrial areas. Taking this into account, we splitted the sample in two groups. The first group contained the policies with sum insured in classes 1 to 6 and the second group contained the policies with sum insured in class 7. For the first group we performed the same modelling strategy as we did in the full sample: A frequency/severity approach and a total loss approach. The outputs of the models for the first group are shown in annexes 3, 4 and 5 and will not be discussed here since most of the tendencies are very similar to the ones described in sections 6.1, 6.2 and 6.4.

For the second group it was empirically observed that the fire cover and the floods cover were responsible for 92.4% (51.8% for the fire cover and 40.6% for the floods cover) of the total claim sizes in this sum insured class. Given the small number of risks in this group we turned to a database with the covers as explanatory factors and the Sum insured as a continous explanatory variable. This database disposal could jeopardize the assumption of independence between risks, since now we cannot be so sure if, e.g., a claim generated by a policy in the fire cover is or isn't correlated with a claim generated by the same policy in the broken glass cover. But since only 13% of the policies were affected by 2 or more claims, that correlation would be small and was ignored. We applied the total loss approach, estimating the p parameter with the algorithm provided in the Tweedie R package (view section 6.4).

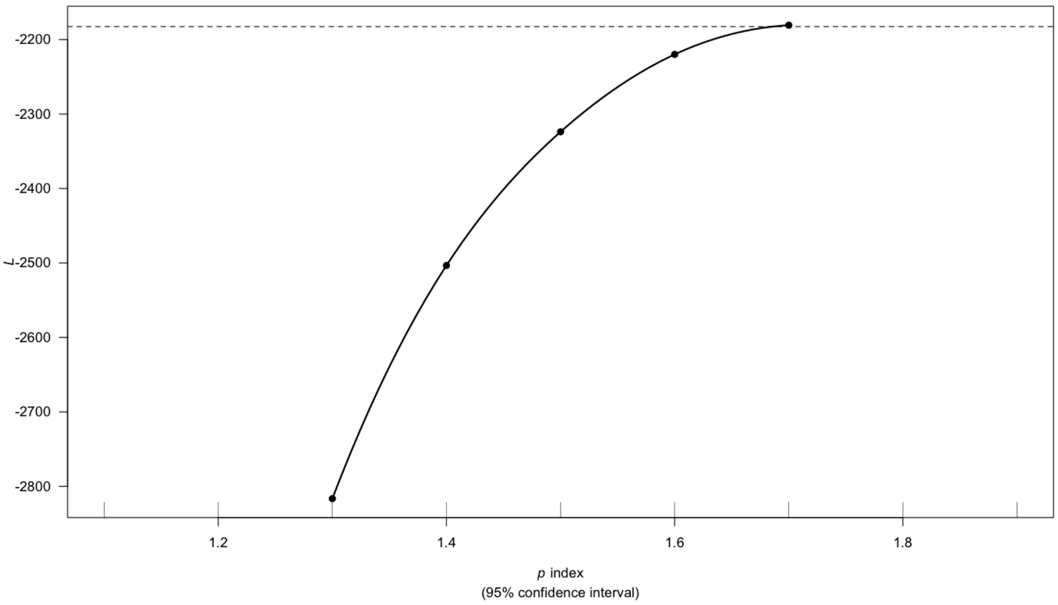


Figure 9: Estimation of the Tweedie p parameter for the sum insured class 7.

The algorithm estimated that a model with p=1.7 will minimize the log-likelihood. However, when we performed the goodness of fit tests applied in sections 6.1 and 6.2, we concluded that the model with p=1.5 was the one which presented the biggest difference between the null deviance and the residual deviance and so this made us quite suspicious of the algorithm results. Estimating the p parameter is a topic out of the scope of this work. It is still subjected to several studies and there are other methods for this estimation from which we highlight the saddlepoint approximation papers by Reid(1988) and Goutis&Casella (1995). This problem will also be addressed in section 6.4.

Still in the second group, we also applied a 2-steps approach to the sample, using again the covers as explanatory factors as the sum insured as a continous explanatory variable. There was no problem in modelling the claim frequency but we struggled with the claim severity model, where we had to make some adjustments, modelling the total claim size (instead of the claim severity) weighted by the number of claims. These models are shown in annexes 7,8 and 9.

6.4 The total loss model

The output for the selected model is shown in annex 6, where the response variable TOTAL_CUSTOS is defined as the total cost of the 2010 claims for each risk. The R package by Peter K Dunn[6] was used in the fitting of the models. The package provides an algorithm that feeds on the model equation, an interval for possible p parameters and a step unit k to move in that interval. It estimates k models and chooses the one with the least log-likelihood. Howhever, there is always the possibility of fitting a model with a given p. For the purpose of this work we are only concerned about the case p∈(1,2) wich characterizes a Poisson-Gamma distribution.

The algorithm didn't converge when assessing our sample, so we divided the interval (1,2) into ten equal intervals and fitted 11 models with the same explanatory variables, one for each p=1, p=1.1, p=1.2, …, p=2, registering the residual deviance in each fitting. The model with the lesser residual deviance would then be used for the forecasting. This is basically what the algorithm does, but with an obvious shorter array of possible p values and substituting the log-likelihood for the residual deviance, hence finding the minimum and not the maximum. The following table summarizes the results we obtained:

[6] http://cran.r-project.org/web/packages/tweedie/

P	Residual Deviance
1	199,019,830
1.1	78,412,998
1.2	32,473,604
1.3	14,196,951
1.4	6,581,006
1.5	3,152,983
1.6	Algorithm did no converge
1.7	Algorithm did no converge
1.8	- Algorithm did no converge
1.9	- Algorithm did no converge
2	- Algorithm did no converge

Table 27: Residual deviance for the tweedie models in respect of the parameter p

For p≥1.6 the algorithm can't fit a model to the data, so we looked at p=1.5 for an experience in forecasting the total IMR loss for 2011. The selected model is shown in Annex 6. The model grasped the behaviour of the Sum Insured factor in an expected way. For the past claim behaviour factor, the model captured the effect that the severity model had also captured, with policies with 3 or more claims in 2010 being less prone to yield claims than policies with 2 claims. Região Autónoma da Madeira, Lezíria do Tejo and Baixo Vouga are again discriminated with the highest risk profiles in the regional factor. In the covers/deductibles factor, the Tweedie model discriminated the variables in a way that was closer to the claim frequency model cover/deductible discrimination.

Performing a similar overall goodness of fit test as the one performed for the other models we get:

(5,237,887-3,152,983=2,084,904) $> \chi^2{}_{(19{,}068-18{,}993=47);95\%} \approx 75$

Which shows again the model is more efficient than the overall mean. The lift chart for this model is also shown in the figure below:

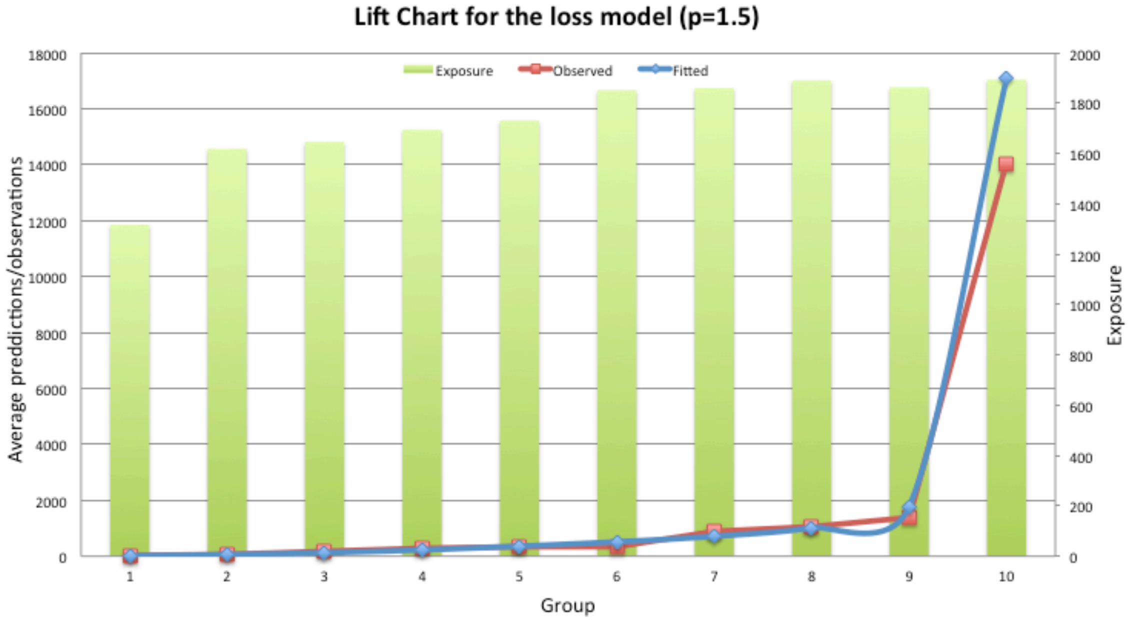


Figure 10: Lift chart for the tweedie model with p=1.5

7 Calculating Best Estimates and dispersion parameters

The objective of all these models is to produce a trusty prediction of future claim payments as well an estimation of their volatility. In this chapter we will start by presenting the aggregate claims expected value and variance, inspired in Centeno, M.L. (2003), fixating the time period in the year 2010 and formulating the problem in the following way,:

- N_i : Number of claims in policy i; i=1,…,n.
- X_{ij} : Claim size i of claim j; for j=1,…, $N_{i.}$

- Y_i : Total claim size of policy i.
- Z : $\sum_{i=1}^{n} Y_i$

Using this formulation and assuming independence in the distributions of N_i, X_i and Y_i , i=1,…n, and between N and X we can calculate the well know result of the expected value of the aggregate claims:

$$E(Z)=E(\sum_{i=1}^{n} Y_i)=\sum_{i=1}^{n} E(Y_i)=\sum_{i=1}^{n} E(N_i)E(X_i)=\sum_{i=1}^{n} \lambda_i\mu_i \tag{9}$$

The variance of the aggregate claims distribution can be calculated in the following way, where the second step of the calculus is possible due to the independence between risks in the r.v. Y:

$$Var(Z)= Var(\sum_{i=1}^{n} Y_i) = \sum_{i=1}^{n} Var(Y_i) = \sum_{i=1}^{n} Var(\sum_{j=1}^{N_i} X_{ij}) = \sum_{i=1}^{n}\{Var[E(\sum_{j=1}^{N_i} X_{ij}|N_i)] + E[Var(\sum_{j=1}^{N_i} X_{ij}|N_i)]\} = \sum_{i=1}^{n}\{[E(X_i)]^2\times Var(N_i) + E(N_i)\times Var(X_i)\}$$

Since $E(N_i) = \lambda_i$, $Var(N_i) = \phi V(\lambda_i) = \phi\lambda_i$, $E(X_i) = \mu_i$ and $Var(X_i) = \sigma_i^2 = \psi\mu_i^2$, we get:

$$Var(Z) = Var(\sum_{i=1}^{n} Y_i) = \sum_{i=1}^{n}(\mu_i^2\phi\lambda_i + \lambda_i\psi\mu_i^2) = (\phi + \psi)\sum_{i=1}^{n} \lambda_i\mu_i^2 \tag{10}$$

If we now take the estimated standard deviation of the compound process $\hat{\sigma}$, i.e., the square root of equation (10), we can calculate a (1-α) risk margin $Z_{\alpha/2}\hat{\sigma}$, where $Z_{\alpha/2}$ represents the Gaussian (0,1) quantile α.

These equations lead us to the following BE's and standard deviations:

Method	Best Estimate	Real 2011 loss	$\hat{\sigma}$	Risk Margin (95%)
Frequency/severity approach (full sample)	**41,749,718€**	**36,894,403€**	**6,878,335€**	**13,481,537€**
Total loss approach (full sample)	**38,807,021€**	**36,894,403€**	**3,487,758€**	**6,836,006€**
A - Frequency/severity approach (split sample – sum insured in classes 1 to 6)	30,696,366€	26,165,641€	5,032,853€	9,864,392€
B - Total loss approach (split sample – sum insured in classes 1 to 6)	27,948,145€	26,165,641€	2,168,741€	4,250,732€
C - Frequency/severity approach (split sample – sum insured in class 7)	9,853,391€	10,728,762€	7,355,506€	14,416,792€
D - Total loss approach (split sample – sum insured in class 7)	13,588,861€	10,728,762€	3,833,150€	7,512,974€
A + C	**40,549,757€**	**36,894,403€**	-	-
A + D	**44,285,227€**	**36,894,403€**	-	-
B + C	**37,801,536€**	**36,894,403€**	-	-
B + D	**38,689,626€**	**36,894,403€**	-	-

Table 28: Summary of the results obtained by the different methods

The estimation of σ didn't took into account the error in the estimation of parameters $\emptyset$ and ψ, so the real $\hat{\sigma}$ would be slightly higher. However, parameter estimation errors are mitigated when we work with large samples and ours was indeed large as table 4 attests. For more on estimating the standard error of the estimated parameter consult England&Verral (1999).

Regarding table 28, if the characteristics of the 2011 sample, i.e, the value of the new explanatory variables were similar to the 2010 sample, we would roughly expect similar claim numbers and amounts, since the annualised exposure was very similar (view table 4). From the lift charts in figures 5 and 6 we would expect a bit of over-estimation, especially from claim frequency side, in the higher risk groups. The slight under-estimation from the claim severity side, again in the higher risk groups, was not enough to balance that over-estimation and the final BE was 13% higher than the real 2011 loss. For the total loss model, the lift chart also showed we could expect over-estimation, as it did happened, even if much lower than in the 2-steps-approach (a 5.2% error). We must also take into account the rough estimate of the Tweedie p parameter with only one decimal place.

The high volatility of the predictions was expected. The economical assimetry of the portuguese industrial tissue seems to be well reflected in these predictions. As figures 7 and 8 show, a small number big portuguese industries are responsible for the sub-branche's greatest part of the losses and even in for the majority of the small/medium sized industries the losses are very volatile. We can attest a higher volatility in the samples with the sum insured in class 7, a lower volatility in the samples with the sum insured in classes 1 to 6, and, as expected, the volatility of the total samples is somewhere in between those two values. The lower volatility of the Tweedie model predictions can be a defining factor if we want to choose one approach, since the BEs are quite similar in both approaches.

When we splitted the sample into more homogeneous groups, the resulting BE's were slightly better. We notice that regardless of the used method, when the BE's of the split models are added, the result is very similar to the BE produced by the full sample model. The worst BE (with a 20% error) comes when we add the BE from the frequency/severity approach for sample with sum insured in classes 1 to 6 with the Tweedie GLM BE for the sum insured in class 7, since both these models over-estimate the 2011 loss.

We must also take into account that the reference period for the measurement of claim numbers and payments was 31.05.2012, wich may leave room for some unsettled claims that would increase the 2011 loss.

8 Conclusions

The MR branch is very important in the portuguese Non-Life segment as well as in all sectors of the portuguese economy. In the last years its premium production surpassed the WC LoB and - given the different social and economical nature of these LoBs - this trend will probably continue in the near future. The strenght of a country's economical and productive tissue is reflected, though not fully of course, in the strenght of the CMR and IMR sub-branches. More industries and more commerce will mean more business for the insurance companies that explore this segment and even in times of economical crisis, as these last 3 years, the premium production of this sub-branch kept growing and the loss ratio was kept stable unlike most of the other Non-Life LoBs.

Internal models are indispensable tools for companies to continuously asses the quality of their business. The Solvency II project has brought more attention to these models and in this work we tried to give a practical example of how to use them in a GLM framework in order to get a better understanding of the IMR sub-branch (or any other LoB), to assess claim payment volatility, to predict claim amounts and therefore to have the ability to calculate technical provisions, SCR's, MCR's and other quantities that are crucial for the management of an insurance company.

Regarding the two approaches we presented, we conclude that the Tweedie GLMs yielded the best predictions in the full sample and in the sample with the sum insured lower than 20,000,000€. Also, it provided a lower standard deviation for the process in all samples. In this approach we only use one model so theoretically this approach will consume half of our time when compared with the frequency/severity approach, but that time will probabily be be spent in accurately estimating the p parameter and several experiments should be made in order to actually get the feel of the Tweedie distribution.

On the other hand, the frequency/severity approach can also be very usefull if we wish to have a grater insight on the variables that influence claim frequency and severity. The modelling strategy to adopt in a problem such as ours will depend on the LoB we are studying, but usually insurance claims show the same behaviour in different LoBs: Many claims with a small size and few claims with a bigger size, with this last group of claims being responsible for the greatest part of the LoB losses. Thus, the segmentation of the sample in homogeneous risk groups will always bring better results. This segmentation can be done by defining a treshold in the claim value or using a variable level to split the sample.

9. References

Antonio, K. & Beirlant J. (2006). Actuarial Statistics With Generalized Linear Mixed Models. *Insurance: Mathematics and economics*, Vol. 40, pp.58-76.

Centeno, M.L. (2003). Teoria do Risco na Actividade Seguradora, 1st Ed. Celta Editora.

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

Dunn, P. (2013). Tweedie: An R Package for Tweedie exponential family models [Online].Available from: http://cran.r-project.org/web/packages/tweedie/.

England, P.D. & Verral, R.J. (1999). Analytic and bootstrap estimates of prediction errors in claim reserving. *Insurance: Mathematics and economics*, Vol. 25, pp.281-293.

Goutis, C. & Casella, G. (1999). Explainig the Saddlepoint approximation. *The American Statistician*, Vol. 53.

Kaas, Rob. 2005. “Compound Poisson Distributions And GLM’s — Tweedie’s Distribution.” Lecture, Royal Flemish Academy of Belgium for Science and the Arts, http://www.kuleuven.be/ucs/seminars_events/other/files/3afm d/Kaas.PDF.

Klugman S., Panjer H. and Willmot G. (2008). Loss Models: From Data to Decisions, 3rd Ed. New York: Wiley.

McCulagh, P. & Nelder, J. (1989). Generalized Linear Models, 2nd Ed. Chapman & Hall/CRC Monographs on Statistics & Applied Probability.

Portuguese Association of Insurers. (2014). “Produção de seguro direto 2013”. [Online].Available from https://segurdata.apseguradores.pt.

Portuguese Association of Insurers. (2014). “Variáveis Trimestrais 2013.12”. [Online].Available from https://segurdata.apseguradores.pt.

Reid, N. Saddlepoint Methods and Statistical Inference. Statistical Science 3 (1988), no. 2, 213--227. doi:10.1214/ss/1177012906. http://projecteuclid.org/euclid.ss/1177012906.

Smyth, G.K., and Jørgensen, B. (2002), Fitting Tweedie’s compound Poisson model to insurance claims data: dispersion modelling, *Astin Bulletin*, pp.143-157.

Tweedie, M.C. (1984). An index that distinguishes between some important exponential families. In Indian Statistical Institute, Statistics: Applications and New Directions, pp.579-604.

Venables W.N and Ripley B.D (2002). Modern Applied Statistics with S, 4th Ed. Springer-Verlag.

Annex 1 – Claim frequency output (full sample)

```
Call:
glm(formula = cbind(TOTAL_CLAIMS) ~ factor(PAST_CLAIM_BEHAVIOUR_1) + 
    factor(PAST_CLAIM_BEHAVIOUR_2) + factor(NUTS3) + factor(CLASS_CAE) + 
    factor(SUM_INSURED_CLASS) + C5 + C8 + C9 + C10 + C12 + C13 + 
    F4 + F5 + F8 + F9 + F10 + F12 + EXPOSURE_YEARS, family = quasipoisson)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-3.3366  -0.5701  -0.3796  -0.2521  14.3656  

Coefficients:
                                        Estimate Std. Error t value Pr(>|t|)    
(Intercept)                             -4.83853    0.42208 -11.463  < 2e-16 ***
factor(PAST_CLAIM_BEHAVIOUR_1)1           0.48466    0.08392   5.775 7.81e-09 ***
factor(PAST_CLAIM_BEHAVIOUR_1)2           0.97291    0.10205   9.534  < 2e-16 ***
factor(PAST_CLAIM_BEHAVIOUR_1)3+          1.66566    0.10135  16.434  < 2e-16 ***
factor(PAST_CLAIM_BEHAVIOUR_1)ausente    -0.12199    0.07032  -1.735  0.08279 .  
factor(PAST_CLAIM_BEHAVIOUR_2)1           0.51873    0.09928   5.225 1.76e-07 ***
factor(PAST_CLAIM_BEHAVIOUR_2)2           0.83310    0.14440   5.769 8.08e-09 ***
factor(PAST_CLAIM_BEHAVIOUR_2)3+          0.74155    0.16113   4.602 4.21e-06 ***
factor(PAST_CLAIM_BEHAVIOUR_2)ausente     0.25276    0.06260   4.038 5.42e-05 ***
factor(NUTS3)ALENTEJO LITORAL            0.75276    0.43923   1.714  0.08657 .  
factor(NUTS3)ALGARVE                     0.02479    0.34228   0.072  0.94226    
factor(NUTS3)ALTO ALENTEJO              -0.67777    0.52841  -1.283  0.19963    
factor(NUTS3)ALTO TRAS-OS-MONTES         0.67591    0.34113   1.981  0.04756 *  
factor(NUTS3)AVE                         0.46842    0.29936   1.565  0.11766    
factor(NUTS3)BAIXO ALENTEJO             -2.06211    1.36114  -1.515  0.12979    
factor(NUTS3)BAIXO MONDEGO               0.29503    0.33221   0.888  0.37450    
factor(NUTS3)BAIXO VOUGA                 0.66727    0.29590   2.255  0.02414 *  
factor(NUTS3)BEIRA INTERIOR NORTE       -0.28876    0.44887  -0.643  0.52004    
factor(NUTS3)BEIRA INTERIOR SUL          0.12723    0.45323   0.281  0.77892    
factor(NUTS3)CAVADO                      0.59772    0.29717   2.011  0.04430 *  
factor(NUTS3)COVA DA BEIRA               0.63503    0.35574   1.785  0.07426 .  
factor(NUTS3)DAO-LAFOES                  0.72057    0.30853   2.335  0.01953 *  
factor(NUTS3)DOURO                       0.49392    0.35637   1.386  0.16577    
factor(NUTS3)ENTRE DOURO E VOUGA         0.44131    0.29808   1.481  0.13875    
factor(NUTS3)GRANDE LISBOA               0.30800    0.29416   1.047  0.29509    
factor(NUTS3)GRANDE PORTO                0.43920    0.29081   1.510  0.13099    
factor(NUTS3)LEZIRIA DO TEJO             0.32375    0.33713   0.960  0.33690    
factor(NUTS3)MEDIO TEJO                 -0.11902    0.35704  -0.333  0.73888    
factor(NUTS3)MINHO-LIMA                  0.52758    0.32557   1.620  0.10514    
factor(NUTS3)NORTE                       0.87636    0.98905   0.886  0.37559    
factor(NUTS3)OESTE                       0.33933    0.31457   1.079  0.28072    
factor(NUTS3)PENINSULA DE SETUBAL       -0.06481    0.33567  -0.193  0.84689    
factor(NUTS3)PINHAL INTERIOR NORTE       0.59535    0.36794   1.618  0.10566    
factor(NUTS3)PINHAL INTERIOR SUL        -0.38431    0.66148  -0.581  0.56126    
factor(NUTS3)PINHAL LITORAL              0.38149    0.31085   1.227  0.21975    
factor(NUTS3)REGIAO AUTONOMA DA MADEIRA  0.92375    0.35249   2.621  0.00878 ** 
factor(NUTS3)REGIAO AUTONOMA DOS ACORES -0.29393    0.43249  -0.680  0.49675    
factor(NUTS3)SERRA DA ESTRELA            1.31494    0.40414   3.254  0.00114 ** 
factor(NUTS3)TAMEGA                      0.50528    0.29637   1.705  0.08823 .  
factor(CLASS_CAE)2                      -0.09834    0.15392  -0.639  0.52288    
factor(CLASS_CAE)3                      -0.48334    0.17045  -2.836  0.00458 ** 
factor(CLASS_CAE)4                      -0.14499    0.13635  -1.063  0.28761    
factor(SUM_INSURED_CLASS)2                 0.29239    0.16878   1.732  0.08323 .  
factor(SUM_INSURED_CLASS)3                 0.52405    0.14137   3.707  0.00021 ***
factor(SUM_INSURED_CLASS)4                 1.06629    0.13306   8.013 1.18e-15 ***
factor(SUM_INSURED_CLASS)5                 1.08898    0.13941   7.811 5.95e-15 ***
factor(SUM_INSURED_CLASS)6                 1.64695    0.13341  12.345  < 2e-16 ***
factor(SUM_INSURED_CLASS)7                 2.21847    0.15221  14.575  < 2e-16 ***
C5                                       1.15922    0.18897   6.134 8.72e-10 ***
C8                                       0.90353    0.17965   5.029 4.96e-07 ***
C9                                      -0.61404    0.18920  -3.245  0.00117 ** 
C10                                      0.31676    0.07595   4.171 3.05e-05 ***
C12                                     -0.33003    0.10991  -3.003  0.00268 ** 
C13                                      0.65102    0.06187  10.522  < 2e-16 ***
F4                                      -0.51006    0.16737  -3.048  0.00231 ** 
F5                                      -0.82172    0.13660  -6.015 1.83e-09 ***
F8                                      -0.61881    0.10767  -5.747 9.20e-09 ***
F9                                       0.45520    0.09173   4.962 7.03e-07 ***
F10                                      0.77362    0.13011   5.946 2.80e-09 ***
F12                                      0.48715    0.11047   4.410 1.04e-05 ***
EXPOSURE_YEARS                           0.74814    0.09765   7.661 1.93e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
(Dispersion parameter for quasipoisson family taken to be 1.779403)

    Null deviance: 16426  on 19068  degrees of freedom
Residual deviance: 12349  on 19008  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 6
```

Annex 2 – Claim severity output (full sample)

```
Call:
glm(formula = cbind(CLAIM_SEVERITY) ~ factor(SUM_INSURED_CLASS) + factor(NUTS3) + 
    factor(CLASS_CAE) + factor(PAST_CLAIM_BEHAVIOUR_1) + C4 + 
    C11 + F5 + F13, family = Gamma(link = "log"), weights = TOTAL_CLAIMS)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-8.1974  -1.9507  -1.2610  -0.2584  11.3420  

Coefficients:
                                          Estimate Std. Error t value Pr(>|t|)    
(Intercept)                              6.5834576  1.0115073   6.509 9.91e-11 ***
factor(SUM_INSURED_CLASS)2                 0.9222684  0.5236370   1.761 0.078369 .  
factor(SUM_INSURED_CLASS)3                 1.3899002  0.4485973   3.098 0.001978 ** 
factor(SUM_INSURED_CLASS)4                 1.4357131  0.4278372   3.356 0.000809 ***
factor(SUM_INSURED_CLASS)5                 1.2917332  0.4414681   2.926 0.003478 ** 
factor(SUM_INSURED_CLASS)6                 1.7691820  0.4264359   4.149 3.51e-05 ***
factor(SUM_INSURED_CLASS)7                 3.1641212  0.4887167   6.474 1.24e-10 ***
factor(NUTS3)ALENTEJO LITORAL            1.4057193  1.2286237   1.144 0.252723    
factor(NUTS3)ALGARVE                     0.9524633  0.9728385   0.979 0.327689    
factor(NUTS3)ALTO ALENTEJO              -1.4418896  1.7216319  -0.838 0.402420    
factor(NUTS3)ALTO TRAS-OS-MONTES        -0.2132291  0.9661720  -0.221 0.825356    
factor(NUTS3)AVE                         0.9458294  0.8421055   1.123 0.261520    
factor(NUTS3)BAIXO ALENTEJO              1.2560231  3.8153473   0.329 0.742042    
factor(NUTS3)BAIXO MONDEGO               1.5534304  0.9527597   1.630 0.103188    
factor(NUTS3)BAIXO VOUGA                 0.4345849  0.8273425   0.525 0.599457    
factor(NUTS3)BEIRA INTERIOR NORTE        2.2071641  1.2504973   1.765 0.077735 .  
factor(NUTS3)BEIRA INTERIOR SUL          0.7439224  1.3179475   0.564 0.572518    
factor(NUTS3)CAVADO                      0.7878922  0.8321778   0.947 0.343881    
factor(NUTS3)COVA DA BEIRA               0.2769778  1.0634488   0.260 0.794546    
factor(NUTS3)DAO-LAFOES                  0.0002461  0.8664153   0.000 0.999773    
factor(NUTS3)DOURO                      -0.5655644  1.0239910  -0.552 0.580805    
factor(NUTS3)ENTRE DOURO E VOUGA         0.5837699  0.8334062   0.700 0.483733    
factor(NUTS3)GRANDE LISBOA               0.9540040  0.8318110   1.147 0.251582    
factor(NUTS3)GRANDE PORTO                0.4318252  0.8131564   0.531 0.595454    
factor(NUTS3)LEZIRIA DO TEJO             3.4402621  1.0152269   3.389 0.000718 ***
factor(NUTS3)MEDIO TEJO                  0.4463964  1.0572956   0.422 0.672927    
factor(NUTS3)MINHO-LIMA                  0.6340641  0.9284734   0.683 0.494755    
factor(NUTS3)NORTE                      -2.6885646  2.7635446  -0.973 0.330755    
factor(NUTS3)OESTE                       0.5039442  0.8778080   0.574 0.565979    
factor(NUTS3)PENINSULA DE SETUBAL        1.0531665  0.9395707   1.121 0.262485    
factor(NUTS3)PINHAL INTERIOR NORTE       0.6174042  1.0354791   0.596 0.551086    
factor(NUTS3)PINHAL INTERIOR SUL         0.0005349  1.8483520   0.000 0.999769    
factor(NUTS3)PINHAL LITORAL              0.2409354  0.8734181   0.276 0.782694    
factor(NUTS3)REGIAO AUTONOMA DA MADEIRA  2.1313460  0.9876827   2.158 0.031071 *  
factor(NUTS3)REGIAO AUTONOMA DOS ACORES  0.1271645  1.3384597   0.095 0.924319    
factor(NUTS3)SERRA DA ESTRELA            0.9166593  1.1284427   0.812 0.416719    
factor(NUTS3)TAMEGA                      0.9968069  0.8278102   1.204 0.228697    
factor(CLASS_CAE)2                       0.3616341  0.5058875   0.715 0.474798    
factor(CLASS_CAE)3                       0.7314827  0.5295058   1.381 0.167321    
factor(CLASS_CAE)4                       1.0178882  0.4475123   2.275 0.023055 *  
factor(PAST_CLAIM_BEHAVIOUR_1)1           0.2982403  0.2408491   1.238 0.215778    
factor(PAST_CLAIM_BEHAVIOUR_1)2           0.5252304  0.2972315   1.767 0.077392 .  
factor(PAST_CLAIM_BEHAVIOUR_1)3+         -0.6042307  0.2817830  -2.144 0.032147 *  
factor(PAST_CLAIM_BEHAVIOUR_1)ausente     0.2004174  0.1792413   1.118 0.263661    
C4                                       0.7126287  0.3697087   1.928 0.054076 .  
C11                                     -0.4584705  0.2643535  -1.734 0.083042 .  
F5                                      -0.3553347  0.1764083  -2.014 0.044135 *  
F13                                     -0.7003232  0.1705725  -4.106 4.22e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 

(Dispersion parameter for Gamma family taken to be 13.83213)

    Null deviance: 10233.8  on 1775  degrees of freedom
Residual deviance:  6919.7  on 1728  degrees of freedom
AIC: 55314

Number of Fisher Scoring iterations: 25
```

Annex 3 – Claim frequency output (Sum insured class in 1 to 6)

```
Call:
glm(formula = cbind(TOTAL_SINISTROS) ~ factor(PAST_CLAIM_BEHAVIOUR_1) + 
    factor(PAST_CLAIM_BEHAVIOUR_2) + factor(NUTS3) + factor(CLASS_CAE) + 
    factor(SUM_INSURED_CLASS) + C2 + C5 + C8 + C10 + C11 + C13 + 
    F2 + F4 + F5 + F7 + F8 + F9 + F10 + F11 + F12 + EXPOSURE_YEARS, 
    family = quasipoisson)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-3.1959  -0.5605  -0.3761  -0.2496  14.0784  

Coefficients:
                                        Estimate Std. Error t value Pr(>|t|)    
(Intercept)                             -5.43273    0.41554 -13.074  < 2e-16 ***
factor(PAST_CLAIM_BEHAVIOUR_1)1           0.49850    0.08553   5.829 5.68e-09 ***
factor(PAST_CLAIM_BEHAVIOUR_1)2           0.88149    0.11236   7.845 4.55e-15 ***
factor(PAST_CLAIM_BEHAVIOUR_1)3+          1.64073    0.11321  14.493  < 2e-16 ***
factor(PAST_CLAIM_BEHAVIOUR_1)ausente    -0.15922    0.07340  -2.169 0.030086 *  
factor(PAST_CLAIM_BEHAVIOUR_2)1           0.43130    0.10355   4.165 3.13e-05 ***
factor(PAST_CLAIM_BEHAVIOUR_2)2           0.92898    0.16069   5.781 7.53e-09 ***
factor(PAST_CLAIM_BEHAVIOUR_2)3+          0.85114    0.18320   4.646 3.41e-06 ***
factor(PAST_CLAIM_BEHAVIOUR_2)ausente     0.22811    0.06476   3.523 0.000428 ***
factor(NUTS3)ALENTEJO LITORAL            0.72627    0.43397   1.674 0.094239 .  
factor(NUTS3)ALGARVE                    -0.09401    0.34602  -0.272 0.785859    
factor(NUTS3)ALTO ALENTEJO              -0.60588    0.52192  -1.161 0.245707    
factor(NUTS3)ALTO TRAS-OS-MONTES         0.61556    0.33813   1.820 0.068700 .  
factor(NUTS3)AVE                         0.40437    0.29680   1.362 0.173070    
factor(NUTS3)BAIXO ALENTEJO             -2.05132    1.34391  -1.526 0.126931    
factor(NUTS3)BAIXO MONDEGO               0.20416    0.33376   0.612 0.540747    
factor(NUTS3)BAIXO VOUGA                 0.67183    0.29305   2.293 0.021887 *  
factor(NUTS3)BEIRA INTERIOR NORTE       -0.18881    0.44332  -0.426 0.670184    
factor(NUTS3)BEIRA INTERIOR SUL         -0.18934    0.57272  -0.331 0.740948    
factor(NUTS3)CAVADO                      0.55006    0.29401   1.871 0.061373 .  
factor(NUTS3)COVA DA BEIRA               0.72373    0.37667   1.921 0.054699 .  
factor(NUTS3)DAO-LAFOES                  0.67408    0.30794   2.189 0.028611 *  
factor(NUTS3)DOURO                       0.43351    0.36124   1.200 0.230137    
factor(NUTS3)ENTRE DOURO E VOUGA         0.40506    0.29536   1.371 0.170260    
factor(NUTS3)GRANDE LISBOA               0.24293    0.29273   0.830 0.406623    
factor(NUTS3)GRANDE PORTO                0.35397    0.28802   1.229 0.219103    
factor(NUTS3)LEZIRIA DO TEJO             0.05201    0.35178   0.148 0.882455    
factor(NUTS3)MEDIO TEJO                 -0.12828    0.35954  -0.357 0.721263    
factor(NUTS3)MINHO-LIMA                  0.44011    0.32458   1.356 0.175141    
factor(NUTS3)NORTE                       0.80482    0.97688   0.824 0.410029    
factor(NUTS3)OESTE                       0.28563    0.31154   0.917 0.359233    
factor(NUTS3)PENINSULA DE SETUBAL       -0.18634    0.34216  -0.545 0.586036    
factor(NUTS3)PINHAL INTERIOR NORTE       0.55909    0.36345   1.538 0.123998    
factor(NUTS3)PINHAL INTERIOR SUL        -0.24830    0.65319  -0.380 0.703849    
factor(NUTS3)PINHAL LITORAL              0.35570    0.30713   1.158 0.246819    
factor(NUTS3)REGIAO AUTONOMA DA MADEIRA  0.94821    0.35282   2.688 0.007205 ** 
factor(NUTS3)REGIAO AUTONOMA DOS ACORES -0.29051    0.44291  -0.656 0.511884    
factor(NUTS3)SERRA DA ESTRELA            1.25654    0.39959   3.145 0.001666 ** 
factor(NUTS3)TAMEGA                      0.46661    0.29279   1.594 0.111021    
factor(CLASS_CAE)2                      -0.15474    0.16169  -0.957 0.338585    
factor(CLASS_CAE)3                      -0.37937    0.18850  -2.013 0.044176 *  
factor(CLASS_CAE)4                      -0.14710    0.14064  -1.046 0.295590    
factor(SUM_INSURED_CLASS)2                 0.32599    0.16710   1.951 0.051087 .  
factor(SUM_INSURED_CLASS)3                 0.56899    0.14032   4.055 5.03e-05 ***
factor(SUM_INSURED_CLASS)4                 1.09815    0.13240   8.294  < 2e-16 ***
factor(SUM_INSURED_CLASS)5                 1.09326    0.13886   7.873 3.65e-15 ***
factor(SUM_INSURED_CLASS)6                 1.64191    0.13424  12.231  < 2e-16 ***
C2                                       0.23531    0.13687   1.719 0.085589 .  
C5                                       0.70364    0.18923   3.718 0.000201 ***
C8                                       0.95469    0.21082   4.528 5.98e-06 ***
C10                                      0.29541    0.08381   3.525 0.000425 ***
C11                                      0.20165    0.10363   1.946 0.051685 .  
C13                                      0.70863    0.06542  10.832  < 2e-16 ***
F2                                      -0.26084    0.15337  -1.701 0.089017 .  
F4                                      -0.41102    0.18325  -2.243 0.024913 *  
F5                                      -0.42742    0.12873  -3.320 0.000901 ***
F7                                       0.24715    0.10983   2.250 0.024445 *  
F8                                      -0.70228    0.14897  -4.714 2.44e-06 ***
F9                                       0.31598    0.09891   3.195 0.001403 ** 
F10                                      0.54691    0.15476   3.534 0.000410 ***
F11                                     -0.16087    0.08240  -1.952 0.050908 .  
F12                                      0.27427    0.09188   2.985 0.002837 ** 
```

```
EXPOSURE_YEARS                           0.66050    0.09820   6.726 1.80e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for quasipoisson family taken to be 1.734842)

    Null deviance: 14946  on 18791  degrees of freedom
Residual deviance: 11761  on 18729  degrees of freedom
  (3 observations deleted due to missingness)
AIC: NA

Number of Fisher Scoring iterations: 6
```

Annex 4 – Claim severity output (Sum insured class in 1 to 6)

```
Call:
glm(formula = cbind(PMTOTAL) ~ factor(SUM_INSURED_CLASS) + factor(NUTS3) + 
    factor(CLASS_CAE) + C11 + C13 + F7, family = Gamma(link = "log"), 
    weights = TOTAL_SINISTROS)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-7.8443  -1.9556  -1.2809  -0.3863  12.2807  

Coefficients:
                                        Estimate Std. Error t value Pr(>|t|)    
(Intercept)                              6.89702    1.10881   6.220 6.27e-10 ***
factor(SUM_INSURED_CLASS)2                 1.32537    0.57318   2.312 0.020883 *
factor(SUM_INSURED_CLASS)3                 1.03751    0.49057   2.115 0.034588 *
factor(SUM_INSURED_CLASS)4                 1.12132    0.46968   2.387 0.017080 *
factor(SUM_INSURED_CLASS)5                 1.20280    0.48024   2.505 0.012355 *
factor(SUM_INSURED_CLASS)6                 1.56822    0.46998   3.337 0.000866 ***
factor(NUTS3)ALENTEJO LITORAL            1.51403    1.34314   1.127 0.259808    
factor(NUTS3)ALGARVE                     0.67652    1.08672   0.623 0.533674    
factor(NUTS3)ALTO ALENTEJO              -1.27781    1.87814  -0.680 0.496374    
factor(NUTS3)ALTO TRAS-OS-MONTES        -0.18903    1.05856  -0.179 0.858298    
factor(NUTS3)AVE                         1.20093    0.92302   1.301 0.193409    
factor(NUTS3)BAIXO ALENTEJO              0.88635    4.17352   0.212 0.831840    
factor(NUTS3)BAIXO MONDEGO               1.77439    1.04320   1.701 0.089148 .  
factor(NUTS3)BAIXO VOUGA                 0.75621    0.90527   0.835 0.403642    
factor(NUTS3)BEIRA INTERIOR NORTE        2.08847    1.36724   1.528 0.126825    
factor(NUTS3)BEIRA INTERIOR SUL          0.60727    1.87836   0.323 0.746510    
factor(NUTS3)CAVADO                      0.80297    0.90892   0.883 0.377130    
factor(NUTS3)COVA DA BEIRA              -0.09550    1.16077  -0.082 0.934439    
factor(NUTS3)DAO-LAFOES                 -0.25005    0.95474  -0.262 0.793425    
factor(NUTS3)DOURO                      -0.90130    1.15379  -0.781 0.434821    
factor(NUTS3)ENTRE DOURO E VOUGA         0.56060    0.91229   0.614 0.538972    
factor(NUTS3)GRANDE LISBOA               0.46323    0.91597   0.506 0.613121    
factor(NUTS3)GRANDE PORTO                0.40575    0.88915   0.456 0.648212    
factor(NUTS3)LEZIRIA DO TEJO             7.86946    1.13158   6.954 5.08e-12 ***
factor(NUTS3)MEDIO TEJO                  0.24036    1.17285   0.205 0.837648    
factor(NUTS3)MINHO-LIMA                  0.75049    1.02589   0.732 0.464547    
factor(NUTS3)NORTE                      -1.82278    3.01122  -0.605 0.545043    
factor(NUTS3)OESTE                       0.35164    0.96003   0.366 0.714203    
factor(NUTS3)PENINSULA DE SETUBAL        1.18799    1.05992   1.121 0.262522    
factor(NUTS3)PINHAL INTERIOR NORTE       0.45783    1.13122   0.405 0.685736    
factor(NUTS3)PINHAL INTERIOR SUL         0.04407    2.02020   0.022 0.982598    
factor(NUTS3)PINHAL LITORAL              0.38368    0.95322   0.403 0.687359    
factor(NUTS3)REGIAO AUTONOMA DA MADEIRA  2.10230    1.09579   1.919 0.055217 .  
factor(NUTS3)REGIAO AUTONOMA DOS ACORES  0.13116    1.46271   0.090 0.928560    
factor(NUTS3)SERRA DA ESTRELA            0.72414    1.22935   0.589 0.555914    
factor(NUTS3)TAMEGA                      0.80233    0.90434   0.887 0.375098    
factor(CLASS_CAE)2                       0.74475    0.55972   1.331 0.183513    
factor(CLASS_CAE)3                       1.01026    0.54444   1.856 0.063690 .  
factor(CLASS_CAE)4                       1.15829    0.48858   2.371 0.017867 *  
C11                                     -0.51359    0.28667  -1.792 0.073391 .  
C13                                     -0.65200    0.20026  -3.256 0.001154 ** 
F7                                      -0.47460    0.18665  -2.543 0.011090 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 

(Dispersion parameter for Gamma family taken to be 16.54649)

    Null deviance: 8290.8  on 1696  degrees of freedom
Residual deviance: 6707.1  on 1655  degrees of freedom
  (3 observations deleted due to missingness)
AIC: 50951

Number of Fisher Scoring iterations: 25
```

Annex 5 – Tweedie GLM output (Sum insured class in 1 to 6)

```
Call:
glm(formula = cbind(TOTAL_CUSTOS) ~ factor(TOTAL_SINISTROS_1_CAT) + 
    ++factor(TOTAL_SINISTROS_2_CAT) + factor(NUTS3) + factor(CLASS_CAPSEGURO) + 
    C1 + C2 + C3 + C4 + C5 + C6 + C7 + C8 + C9 + C10 + C11 + 
    C12 + C13 + C15 + +F1 + F2 + F3 + F4 + F5 + F6 + F7 + F8 + 
    F9 + F10 + F11 + F12 + F13 + F15 + EXPOSURE_YEARS, family = tweedie(var.power = 1.6, 
    link.power = 0))

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-27.215   -8.563   -6.848   -5.069  145.934  

Coefficients: (1 not defined because of singularities)
                                         Estimate Std. Error t value Pr(>|t|)    
(Intercept)                             -29.22281  812.83645  -0.036 0.971321    
factor(TOTAL_SINISTROS_1_CAT)1            0.64575    0.27829   2.320 0.020332 *  
factor(TOTAL_SINISTROS_1_CAT)2            1.42940    0.40207   3.555 0.000379 ***
factor(TOTAL_SINISTROS_1_CAT)3+           1.41717    0.58562   2.420 0.015532 *  
factor(TOTAL_SINISTROS_1_CAT)ausente     -0.07666    0.21141  -0.363 0.716891    
factor(TOTAL_SINISTROS_2_CAT)1            0.65587    0.34528   1.900 0.057511 .  
factor(TOTAL_SINISTROS_2_CAT)2            1.79424    0.56978   3.149 0.001641 ** 
factor(TOTAL_SINISTROS_2_CAT)3+           0.23544    1.00614   0.234 0.814981    
factor(TOTAL_SINISTROS_2_CAT)ausente      0.20534    0.18691   1.099 0.271947    
factor(NUTS3)ALENTEJO LITORAL             1.15234    1.30978   0.880 0.378981    
factor(NUTS3)ALGARVE                      0.86478    0.95357   0.907 0.364477    
factor(NUTS3)ALTO ALENTEJO               -2.08207    1.69432  -1.229 0.219143    
factor(NUTS3)ALTO TRAS-OS-MONTES          0.52668    1.09394   0.481 0.630200    
factor(NUTS3)AVE                          1.73600    0.86494   2.007 0.044755 *  
factor(NUTS3)BAIXO ALENTEJO              -1.99968    2.13171  -0.938 0.348224    
factor(NUTS3)BAIXO MONDEGO                1.44853    0.94247   1.537 0.124323    
factor(NUTS3)BAIXO VOUGA                  1.53644    0.86888   1.768 0.077029 .  
factor(NUTS3)BEIRA INTERIOR NORTE         1.33232    1.16760   1.141 0.253854    
factor(NUTS3)BEIRA INTERIOR SUL          -0.18198    1.64144  -0.111 0.911724    
factor(NUTS3)CAVADO                       1.60143    0.86731   1.846 0.064846 .  
factor(NUTS3)COVA DA BEIRA                1.44240    1.19194   1.210 0.226245    
factor(NUTS3)DAO-LAFOES                   0.90923    0.94763   0.959 0.337329    
factor(NUTS3)DOURO                       -0.28510    1.22635  -0.232 0.816172    
factor(NUTS3)ENTRE DOURO E VOUGA          1.21641    0.86925   1.399 0.161719    
factor(NUTS3)GRANDE LISBOA                0.50059    0.85810   0.583 0.559653    
factor(NUTS3)GRANDE PORTO                 1.15790    0.84485   1.371 0.170535    
factor(NUTS3)LEZIRIA DO TEJO              3.67701    0.88163   4.171 3.05e-05 ***
factor(NUTS3)MEDIO TEJO                   0.98234    0.98550   0.997 0.318876    
factor(NUTS3)MINHO-LIMA                   0.57269    1.00577   0.569 0.569089    
factor(NUTS3)NORTE                       -0.70423    5.48416  -0.128 0.897825    
factor(NUTS3)OESTE                        0.98326    0.91520   1.074 0.282670    
factor(NUTS3)PENINSULA DE SETUBAL         1.08694    0.91932   1.182 0.237089    
factor(NUTS3)PINHAL INTERIOR NORTE        1.28010    1.07756   1.188 0.234864    
factor(NUTS3)PINHAL INTERIOR SUL          0.07643    1.71427   0.045 0.964441    
factor(NUTS3)PINHAL LITORAL               0.99523    0.90969   1.094 0.273953    
factor(NUTS3)REGIAO AUTONOMA DA MADEIRA   3.05000    0.97195   3.138 0.001704 ** 
factor(NUTS3)REGIAO AUTONOMA DOS ACORES  -0.57532    1.32628  -0.434 0.664450    
factor(NUTS3)SERRA DA ESTRELA             3.73111    1.18680   3.144 0.001670 ** 
factor(NUTS3)TAMEGA                       1.49208    0.85943   1.736 0.082558 .  
factor(CLASS_CAPSEGURO)2                  1.41496    0.42283   3.346 0.000820 ***
factor(CLASS_CAPSEGURO)3                  2.08791    0.36472   5.725 1.05e-08 ***
factor(CLASS_CAPSEGURO)4                  2.78601    0.35989   7.741 1.03e-14 ***
factor(CLASS_CAPSEGURO)5                  2.78105    0.38946   7.141 9.62e-13 ***
factor(CLASS_CAPSEGURO)6                  3.60961    0.38755   9.314  < 2e-16 ***
C1                                       -0.34700    0.48138  -0.721 0.471023    
C2                                        0.33512    0.43659   0.768 0.442747    
C3                                       -2.13089    1.85124  -1.151 0.249724    
C4                                       -2.02446    2.50272  -0.809 0.418581    
C5                                        1.54193    0.55897   2.759 0.005812 ** 
C6                                        0.01590    0.25940   0.061 0.951117    
C7                                       28.36927  812.83604   0.035 0.972159    
C8                                        2.81498    2.03219   1.385 0.166009    
C9                                        1.43223    0.77835   1.840 0.065771 .  
C10                                      -0.21834    0.27659  -0.789 0.429883    
C11                                      -0.61841    0.29606  -2.089 0.036740 *  
C12                                      -0.20446    0.38448  -0.532 0.594879    
C13                                       0.63433    0.36360   1.745 0.081078 .  
C15                                      -0.03961    1.19223  -0.033 0.973499    
F1                                        0.69430    0.51331   1.353 0.176202    
F2                                       -0.77564    0.46517  -1.667 0.095445 .  
F3                                        0.54937    1.94753   0.282 0.777880    
```

```
F4                                      2.07047    2.54443   0.814 0.415812    
F5                                     -1.30450    0.44204  -2.951 0.003171 ** 
F6                                           NA         NA      NA       NA    
F7                                      0.36149    0.32760   1.103 0.269852    
F8                                     -1.99742    1.99479  -1.001 0.316686    
F9                                      0.29003    0.31131   0.932 0.351534    
F10                                     1.00478    0.52154   1.927 0.054051 .  
F11                                     0.47882    0.25487   1.879 0.060299 .  
F12                                     0.20675    0.38513   0.537 0.591383    
F13                                    -0.57361    0.38866  -1.476 0.139990    
F15                                     0.42454    0.61229   0.693 0.488090    
EXPOSURE_YEARS                          1.99820    0.29843   6.696 2.21e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Tweedie family taken to be 1190.243)

    Null deviance: 2196168  on 18791  degrees of freedom
Residual deviance: 1565215  on 18720  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 22
```

Annex 6 – Tweedie GLM output (full sample)

```
Call:
glm(formula = cbind(TOTAL_CUSTOS) ~ factor(PAST_CLAIM_BEHAVIOUR_1) + 
    factor(PAST_CLAIM_BEHAVIOUR_2) + factor(NUTS3) + factor(CLASS_CAE) + 
    factor(SUM_INSURED_CLASS) + C1 + C2 + C3 + C4 + C5 + C6 + C7 + 
    C8 + C9 + C10 + C11 + C12 + C13 + C15 + F1 + F2 + F3 + F4 + 
    F5 + F7 + F8 + F9 + F10 + F11 + F12 + F13 + F15 + EXPOSURE_YEARS, 
    family = tweedie(var.power = 1.5, link.power = 0))

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-75.585  -10.846   -8.413   -5.937  238.700  

Coefficients:
                                         Estimate Std. Error t value Pr(>|t|)    
(Intercept)                             -23.78954  606.81068  -0.039 0.968728    
factor(PAST_CLAIM_BEHAVIOUR_1)1           0.62705    0.25521   2.457 0.014020 *  
factor(PAST_CLAIM_BEHAVIOUR_1)2           1.85465    0.29726   6.239 4.49e-10 ***
factor(PAST_CLAIM_BEHAVIOUR_1)3+          1.11604    0.46470   2.402 0.016332 *  
factor(PAST_CLAIM_BEHAVIOUR_1)ausente    -0.01165    0.19451  -0.060 0.952225    
factor(PAST_CLAIM_BEHAVIOUR_2)1           0.68949    0.31188   2.211 0.027066 *  
factor(PAST_CLAIM_BEHAVIOUR_2)2           1.42051    0.47643   2.982 0.002871 ** 
factor(PAST_CLAIM_BEHAVIOUR_2)3+          0.56903    0.68511   0.831 0.406229    
factor(PAST_CLAIM_BEHAVIOUR_2)ausente     0.43890    0.17217   2.549 0.010803 *  
factor(NUTS3)ALENTEJO LITORAL            1.45036    1.24053   1.169 0.242358    
factor(NUTS3)ALGARVE                     0.85062    0.94865   0.897 0.369908    
factor(NUTS3)ALTO ALENTEJO              -2.42348    1.85706  -1.305 0.191904    
factor(NUTS3)ALTO TRAS-OS-MONTES         0.32482    1.11744   0.291 0.771296    
factor(NUTS3)AVE                         1.51747    0.86674   1.751 0.079999 .  
factor(NUTS3)BAIXO ALENTEJO             -1.62812    2.13820  -0.761 0.446402    
factor(NUTS3)BAIXO MONDEGO               1.17233    0.94635   1.239 0.215435    
factor(NUTS3)BAIXO VOUGA                 1.69802    0.86476   1.964 0.049594 *  
factor(NUTS3)BEIRA INTERIOR NORTE        1.52560    1.10265   1.384 0.166504    
factor(NUTS3)BEIRA INTERIOR SUL          0.71557    1.30780   0.547 0.584280    
factor(NUTS3)CAVADO                      1.42622    0.87071   1.638 0.101438    
factor(NUTS3)COVA DA BEIRA               1.19555    1.16187   1.029 0.303499    
factor(NUTS3)DAO-LAFOES                  0.84787    0.94021   0.902 0.367182    
factor(NUTS3)DOURO                      -0.31861    1.23459  -0.258 0.796353    
factor(NUTS3)ENTRE DOURO E VOUGA         1.20193    0.86952   1.382 0.166895    
factor(NUTS3)GRANDE LISBOA               0.79038    0.85597   0.923 0.355824    
factor(NUTS3)GRANDE PORTO                1.04947    0.84968   1.235 0.216796    
factor(NUTS3)LEZIRIA DO TEJO             3.32394    0.87680   3.791 0.000151 ***
factor(NUTS3)MEDIO TEJO                  0.60061    0.99110   0.606 0.544520    
factor(NUTS3)MINHO-LIMA                  0.54407    0.99610   0.546 0.584938    
factor(NUTS3)NORTE                      -0.54226    5.79207  -0.094 0.925411    
factor(NUTS3)OESTE                       0.84675    0.92161   0.919 0.358224    
factor(NUTS3)PENINSULA DE SETUBAL        0.92006    0.91595   1.004 0.315159    
factor(NUTS3)PINHAL INTERIOR NORTE       1.13682    1.08007   1.053 0.292560    
factor(NUTS3)PINHAL INTERIOR SUL        -0.24113    1.72873  -0.139 0.889068    
factor(NUTS3)PINHAL LITORAL              0.91019    0.91210   0.998 0.318339    
factor(NUTS3)REGIAO AUTONOMA DA MADEIRA  2.77465    0.94578   2.934 0.003353 ** 
factor(NUTS3)REGIAO AUTONOMA DOS ACORES -0.82453    1.33894  -0.616 0.538030    
factor(NUTS3)SERRA DA ESTRELA            3.48716    1.11801   3.119 0.001817 ** 
factor(NUTS3)TAMEGA                      1.40874    0.86366   1.631 0.102878    
factor(CLASS_CAE)2                       0.88106    0.63092   1.396 0.162594    
factor(CLASS_CAE)3                       1.31706    0.65201   2.020 0.043396 *  
factor(CLASS_CAE)4                       1.47984    0.58206   2.542 0.011017 *  
factor(SUM_INSURED_CLASS)2                 1.73073    0.43913   3.941 8.13e-05 ***
factor(SUM_INSURED_CLASS)3                 1.99302    0.39353   5.065 4.13e-07 ***
factor(SUM_INSURED_CLASS)4                 2.74861    0.38523   7.135 1.00e-12 ***
factor(SUM_INSURED_CLASS)5                 2.73263    0.40916   6.679 2.48e-11 ***
factor(SUM_INSURED_CLASS)6                 3.49642    0.40446   8.645  < 2e-16 ***
factor(SUM_INSURED_CLASS)7                 5.35276    0.44394  12.057  < 2e-16 ***
C1                                      -1.04240    0.39304  -2.652 0.008004 ** 
C2                                       0.64700    0.37537   1.724 0.084795 .  
C3                                      -1.80997    1.78250  -1.015 0.309922    
C4                                      -1.43699    2.44912  -0.587 0.557386    
C5                                       1.38803    0.52161   2.661 0.007797 ** 
C6                                       0.02815    0.24182   0.116 0.907327    
C7                                      21.48733  606.80991   0.035 0.971753    
C8                                       2.46303    1.92784   1.278 0.201401    
C9                                       1.47091    0.70802   2.077 0.037769 *  
C10                                     -0.28114    0.25937  -1.084 0.278407    
C11                                     -0.36390    0.26565  -1.370 0.170742    
C12                                     -0.27147    0.35167  -0.772 0.440151    
C13                                      0.55907    0.31975   1.748 0.080406 .  
```

```
C15                                         0.19556    1.15073   0.170 0.865052    
F1                                          1.27905    0.42636   3.000 0.002704 ** 
F2                                         -1.11343    0.40512  -2.748 0.005995 ** 
F3                                          0.55799    1.85545   0.301 0.763624    
F4                                          1.38019    2.49447   0.553 0.580063    
F5                                         -1.10048    0.40924  -2.689 0.007171 ** 
F7                                          0.51545    0.31617   1.630 0.103060    
F8                                         -1.55645    1.88858  -0.824 0.409871    
F9                                          0.36803    0.29868   1.232 0.217900    
F10                                         1.45610    0.45254   3.218 0.001295 ** 
F11                                         0.09616    0.22774   0.422 0.672849    
F12                                         0.46176    0.36151   1.277 0.201507    
F13                                        -0.66740    0.34506  -1.934 0.053107 .  
F15                                        -0.09416    0.60972  -0.154 0.877270    
EXPOSURE_YEARS                              1.87690    0.28716   6.536 6.47e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Tweedie family taken to be 2294.201)

    Null deviance: 5237887  on 19068  degrees of freedom
Residual deviance: 3152983  on 18993  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 15
```

Annex 7 – Tweedie GLM output (Sum insured in class 7)

```
Call:
glm(formula = cbind(valor) ~ factor(cobertura) + CAPSEGURO, family = tweedie(var.power =
1.5,
    link.power = 0))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-96.264  -13.053   -9.926   -4.663  195.198

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)          7.195e+00  6.578e-01  10.938  < 2e-16 ***
factor(cobertura)2  -3.177e+01  1.675e+03  -0.019  0.98487
factor(cobertura)3  -1.637e+00  1.027e+00  -1.594  0.11100
factor(cobertura)4  -3.282e+01  1.627e+04  -0.002  0.99839
factor(cobertura)5  -9.318e-01  9.272e-01  -1.005  0.31507
factor(cobertura)6  -3.175e+01  1.447e+03  -0.022  0.98250
factor(cobertura)7   2.320e+00  7.134e-01   3.252  0.00117 **
factor(cobertura)8   1.435e+00  7.493e-01   1.915  0.05563 .
factor(cobertura)9  -3.789e-01  8.924e-01  -0.425  0.67116
factor(cobertura)10 -3.176e+01  3.938e+03  -0.008  0.99357
factor(cobertura)11 -3.954e+00  1.701e+00  -2.324  0.02023 *
factor(cobertura)12 -3.945e+00  2.204e+00  -1.790  0.07366 .
factor(cobertura)13 -1.156e+00  1.279e+00  -0.904  0.36607
factor(cobertura)15  1.392e-01  8.186e-01   0.170  0.86502
CAPSEGURO            4.295e-09  4.027e-10  10.666  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Tweedie family taken to be 2304.453)

    Null deviance: 1163343  on 1914  degrees of freedom
Residual deviance:  518086  on 1900  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 22
```

Annex 8 – Claim frequency output (Sum insured in class 7)

```
Call:
glm(formula = cbind(NS) ~ factor(Cobertura) + CAPSEGURO, family = quasipoisson)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6350  -0.6596  -0.4911  -0.2604   6.9368

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)        -2.344e+00  3.782e-01  -6.197 7.03e-10 ***
factor(COVER)2  -1.609e+01  8.301e+02  -0.019   0.9845
factor(COVER)3   1.443e-01  4.686e-01   0.308   0.7581
factor(COVER)4  -1.661e+01  8.080e+03  -0.002   0.9984
factor(COVER)5   7.965e-01  4.266e-01   1.867   0.0620 .
factor(COVER)6  -1.608e+01  7.176e+02  -0.022   0.9821
factor(COVER)7   6.578e-01  4.294e-01   1.532   0.1257
factor(COVER)8  -8.109e-01  5.736e-01  -1.414   0.1576
factor(COVER)9   4.452e-01  4.529e-01   0.983   0.3257
factor(COVER)10 -1.608e+01  1.951e+03  -0.008   0.9934
factor(COVER)11 -1.116e+00  6.896e-01  -1.618   0.1057
factor(COVER)12  2.003e-01  5.560e-01   0.360   0.7187
factor(COVER)13  1.174e+00  4.599e-01   2.553   0.0108 *
factor(COVER)15  8.362e-01  4.224e-01   1.980   0.0479 *
SUM_INSURED      2.120e-09  3.017e-10   7.025 2.96e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for quasipoisson family taken to be 1.997221)

    Null deviance: 1573.5  on 1918  degrees of freedom
Residual deviance: 1356.5  on 1904  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 16
```

Annex 9 – Claim severity output (Sum insured in class 7)

```
Call:
glm(formula = cbind(Valor) ~ factor(Cobertura) + CAPSEGURO, family = Gamma(link = "log"),
    weights = NS)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-10.9818   -2.4474   -1.6756   -0.1854    5.1589

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)          9.624e+00  6.146e-01  15.659  < 2e-16 ***
factor(COVER)3  -1.317e+00  7.616e-01  -1.729 0.085946 .
factor(COVER)5  -3.965e-01  6.941e-01  -0.571 0.568717
factor(COVER)7   2.655e+00  6.988e-01   3.799 0.000214 ***
factor(COVER)8   2.706e+01  9.675e-01  27.963  < 2e-16 ***
factor(COVER)9   2.113e-02  7.361e-01   0.029 0.977136
factor(COVER)11 -2.804e+00  1.121e+00  -2.501 0.013492 *
factor(COVER)12 -3.222e+00  9.037e-01  -3.565 0.000494 ***
factor(COVER)13 -1.278e+00  7.476e-01  -1.710 0.089456 .
factor(COVER)15 -4.171e-01  6.866e-01  -0.607 0.544494
SUM_INSURED      9.513e-10  5.582e-10   1.704 0.090477 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Gamma family taken to be 5.27641)

    Null deviance: 1687.0  on 154  degrees of freedom
Residual deviance: 1226.2  on 144  degrees of freedom
AIC: 6372

Number of Fisher Scoring iterations: 25
```