Pricing of Car Insurance
with Generalized Linear Models

door

Evelien Brisard

Promotor Prof. Robert Verlaak en Begeleider Ellen Van den Acker

Manamaproef ingediend tot het behalen van master-na-master in de
Actuariële wetenschappen

Academiejaar 2013–2014
Pricing of Car Insurance
with Generalized Linear Models

door

Evelien Brisard

Promotor Prof. Robert Verlaak en Begeleider Ellen Van den Acker

Manamaproef ingediend tot het behalen van master-na-master in de Actuariële wetenschappen

Academiejaar 2013–2014
Preface

A lot of insurers use software provided by specialized companies. Often this software is more user-friendly and easier to interpret than standard statistical software, but the problem is that the users often do not know what is behind the numbers and conclusions. To solve this black-box problem, a lot of knowledge and know-how has to be created by research, practice and experience with statistical software and the data itself. I do not have the experience, but I tried to capture the most important notions concerning generalized linear models through research. The practice in my case was on a real dataset and required lots of hours, days, weeks of trying and searching in SAS. It was not easy and definitely the last mile is the longest, but I did it, hereby I thank my family and friends to keep supporting me. I am also grateful for the opportunity to link this thesis to my work. These last two years of actuarial science were very interesting and the internship made the perfect transition to the beginning of my career. I hope you enjoy my thesis!

Evelien Brisard, may 2014
Authors’ rights

All rights reserved. Nothing from this thesis can be copied, used, quoted, further studied or made public, nor electronically, nor manually, nor in any other way, without the explicit written consent of the writer, Evelien Brisard.

Evelien Brisard, may 2014
Summary

Third party liability car insurance is one of the most important insurances in many countries, as it is obligatory and thus lots of data are disposable. The tarification however is often a difficult exercise since different explanatory variables are available and often a long history precedes the analysis. Generalized linear models can be a great way to efficiently predict important ratios, like the claim frequency, claim severity and pure premium. In this thesis I choose to work with SAS because it handles large datasets very well and it was available to me; however also other statistical programs have a number of tools to study GLM. In a first part of this thesis, the theory of GLM is explained. The reader gains insight in different distributions and their properties, various ways to build a model and the meaning of the given output or outcomes. In the second part, this is then applied to a realistic dataset. The variables are introduced and discussed in chapter four, both a priori and a posteriori explanatory variables, the response variables, and possible dependencies are studied. A way to determine a segmentation of continuous variables is shown and the importance and possibilities of using interactions are discussed. Then in chapter six, through forward (stepwise) regression a model is build, both a frequency and severity model, leading to a predicted premium. This is furthermore compared with the predicted pure premium from a Tweedie model, and with the original earned premium. Other distribution or modelling possibilities are briefly discussed in the final chapter.
Keywords

car insurance, insurance pricing, generalized linear models, GLM, SAS, pricing modelling, pure premium modelling, Poisson model, Gamma model, Tweedie models, segmentation, predictive modelling, GENMOD statement, exposure
Contents

Preface ii
Authors’ rights iii
Summary iv
Table of contents vi
Used abbreviations ix

1 General introduction 1
  1.1 Risk classification ............................................. 2
  1.2 Bonus Malus systems ........................................... 4
  1.3 Exposure, claim frequency and severity, pure premium ................. 6
      1.3.1 Assumptions ............................................. 7
      1.3.2 Properties .............................................. 7
  1.4 Claim counts .................................................... 9
      1.4.1 Poisson distribution .................................... 9
      1.4.2 Mixed Poisson distribution .............................. 10
      1.4.3 Compound Poisson distribution ......................... 12
  1.5 Claim severity .................................................. 12

2 Generalized Linear Models 14
  2.1 General model .................................................. 15
  2.2 Estimators ...................................................... 17
  2.3 Confidence interval ............................................ 18
  2.4 Estimation of φ ............................................... 19
  2.5 Poisson model .................................................. 20
  2.6 Gamma distribution ........................................... 22
  2.7 Multiplicative tariff .......................................... 22
  2.8 Offset .......................................................... 23
## CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.9 Weight</td>
<td>24</td>
</tr>
<tr>
<td>2.10 Tariff</td>
<td>26</td>
</tr>
<tr>
<td>2.11 Tweedie models</td>
<td>26</td>
</tr>
<tr>
<td>2.12 Goodness of fit measures</td>
<td>30</td>
</tr>
<tr>
<td>3 Credibility theory</td>
<td>32</td>
</tr>
<tr>
<td>3.1 Credibility estimator</td>
<td>32</td>
</tr>
<tr>
<td>3.2 Bülmann-Straub model</td>
<td>33</td>
</tr>
<tr>
<td>3.3 Combine GLM with credibility theory</td>
<td>35</td>
</tr>
<tr>
<td>4 Dataset</td>
<td>36</td>
</tr>
<tr>
<td>4.1 Variables: overview</td>
<td>36</td>
</tr>
<tr>
<td>4.1.1 A priori variables</td>
<td>36</td>
</tr>
<tr>
<td>4.1.2 A posteriori variables</td>
<td>38</td>
</tr>
<tr>
<td>4.2 Univariate statistics and first insights</td>
<td>39</td>
</tr>
<tr>
<td>4.2.1 Overall view</td>
<td>41</td>
</tr>
<tr>
<td>4.2.2 Age of the vehicle, policy and policyholder</td>
<td>43</td>
</tr>
<tr>
<td>4.2.3 Geographical zones</td>
<td>47</td>
</tr>
<tr>
<td>4.2.4 Vehicle characteristics</td>
<td>53</td>
</tr>
<tr>
<td>4.2.5 Policy characteristics</td>
<td>55</td>
</tr>
<tr>
<td>4.2.6 History of the policyholder and other</td>
<td>57</td>
</tr>
<tr>
<td>4.3 Dependencies between the explanatory variables</td>
<td>61</td>
</tr>
<tr>
<td>5 Segmentation and interaction</td>
<td>70</td>
</tr>
<tr>
<td>5.1 Variables and segmentation</td>
<td>71</td>
</tr>
<tr>
<td>5.1.1 Segmentation of the age of the policyholder</td>
<td>72</td>
</tr>
<tr>
<td>5.1.2 Segmentation of BM</td>
<td>75</td>
</tr>
<tr>
<td>5.1.3 Segmentation of power</td>
<td>76</td>
</tr>
<tr>
<td>5.1.4 Segmentation of age_car</td>
<td>78</td>
</tr>
<tr>
<td>5.1.5 Linearity of newzone</td>
<td>79</td>
</tr>
<tr>
<td>5.2 Interactions of explanatory variables</td>
<td>80</td>
</tr>
<tr>
<td>5.2.1 Two categorical variables</td>
<td>80</td>
</tr>
<tr>
<td>5.2.2 Three categorical variables</td>
<td>85</td>
</tr>
<tr>
<td>5.3 Application: BS estimators for zoneG</td>
<td>85</td>
</tr>
<tr>
<td>6 Building the model</td>
<td>87</td>
</tr>
<tr>
<td>6.1 Assumptions and preconditions</td>
<td>87</td>
</tr>
<tr>
<td>6.1.1 Explanatory variables</td>
<td>87</td>
</tr>
<tr>
<td>6.1.2 Method</td>
<td>88</td>
</tr>
<tr>
<td>6.2 Frequency model selection</td>
<td>89</td>
</tr>
</tbody>
</table>
Used abbreviations

\( \overset{d}{=} \) is distributed as
AIC Akaike information criterion
BE Best Estimate
BM Bonus Malus
BS Büllmann-Straub
cgf cumulant generating function \( (\Psi_Y(z) = \ln M_Y(z)) \)
CI Confidence Interval
GLM Generalized Linear Models
LR Loss Ratio
mgf moment generating function \( (M_Y(z) = E[\exp(zY)]) \)
ML Maximum likelihood
MLE Maximum likelihood estimator
MLF Multi-level factor
pgf probability generating function \( (\phi_Y(z) = E[z^Y]) \)
TPL Third party liability
Chapter 1

General introduction

All kinds of accidents happen in our daily lives where somebody is responsible for the damage, however in most cases the harm was not done on purpose. Think about traffic accidents, fires, incidents resulting from lack of guidance or faults in construction, and so on. For the parties involved, this would be a tremendous amount to repay, in most cases leading to debts for the rest of their lives, if they were not insured.

After an incident where the insured person or family or business becomes responsible for the damage caused to someone or something else, the insurer, to which premiums were paid by the insured, has to reimburse the damage. This of course only up to certain (legal) limits within the boundaries stated in the insurance contract. These losses are compensated by the premiums earned by the insurer, which have to be paid (mostly on an annual basis) by all clients, regardless whether there is a claim that year or not (an insurance contract is a risk contract). So the number of clients that experience no claims or liability, pay for the number of clients that become liable for certain damage (principle of insurance - dispersion of risk). Hence in order to balance the lossratio, the amount of loss from claims that have to be paid out divided by the total premium income, not only the premiums have to be determined carefully, but also the clients have to be chosen wisely. Since not everybody has the same probability to catch a claim: a house with a skeleton structure of wood in stead of steel, will burn down more likely; estate in a country that experiences many earthquakes will produce more claims than estate in Belgium in terms of a contract that insures against fire and natural disasters.

The economic risk is transferred from the policyholder to the insurer, and this works because of the law of large numbers. The insurer has a large number of similar policies so that
1.1 Risk classification

his portfolio of risks becomes more predictable and behaves like the expected value of the portfolio. The goal is always to maintain a healthy loss/profit ratio on their balance sheet, and this by reducing the variability around this expected value - the volatility. By applying one tariff for all policyholders, there will be a lot of volatility in the insurer’s portfolio since not all the contracts are evenly risky. Moreover, better clients will feel neglected since they have to pay the same premium, and will go to competitive insurers. This leads to adverse selection where the good, more profitable clients will leave the insurer who will be left with underpriced riskier contracts. A differentiated tariff is the solution, with different premiums for different risk categories. Different categories have different probabilities to produce claims, so it is extremely important to choose these categories wisely. Adverse selection is limited and the volatility is reduced since the expected values are adapted to the different risk levels.  

In this thesis, I will discuss car (or motor) insurance. The coverage can be divided into first and third party coverage: the first party coverage protects the vehicle owner in case he is responsible for the accident, where third party coverage protects other parties involved that were not responsible. Note that not responsible not necessarily means that this third party had absolutely no fault: think about the law concerning vulnerable road users. This applies to passengers on foot, cyclists, non-driving occupants of the vehicle, and states that these are reimbursed, regardless their rate of responsibility in the incident.

In Belgium, as in most countries, a third party liability (TPL) coverage is required to be allowed on the public road. This makes that in non-life insurance, car insurance represents a large share of the policies, and can maybe even be called the core of many (non-life) insurers’ business. Moreover there is extensive information available about the characteristics of the policyholders, and this all explains the devoted research and developed methods for (third party liability) car insurance. In the dataset used here, only accidents where the represented policyholder was in fault are considered since otherwise, other insurers will had to pay for the harm done.

1.1 Risk classification

To obtain different risk categories, one uses variables to divide the policyholders. A priori variables are variables which values can be determined before the policyholders start to drive. Individual characteristics are variables that describe the policyholder (note that this
1.1 Risk classification

may not be the future driver!) like his age, gender, .... (Note that the gender is not longer allowed as a tariff indicator by European law.) Motor characteristics are variables that describe the insured vehicle like the age of the vehicle, fuel type, use of the vehicle, .... Geographical characteristics describe the living area or environment of the policyholder. Note that most variables are categoric variables, meaning that they can take a number of values that represent the levels but have no further continuous meaning: if 0 means professional use and 1 private use, one can change these into 1 and 2 or a and b but this doesn’t change the analysis. The generalized regression models associate a parameter to each level separately. But age can be used as a continuous variable where the model then associates only one parameter that reflects the change in the variable if it increases by one (continuously). However seldom a linear relationship is observed and moreover, categorical variables are much more compatible with the risk classification and use of tariff cells. For the classification variables in the analysis, we will always choose the level with the largest amount of exposure as base level and will hence produce the base rate. This will become clear when models will be written out later.

These a priori variables results in an unfair rating system since no correction is made for evidence of good or bad driving skills. **Experience rating** uses the claim history of the individual in the form of *a posteriori* variables to adjust the premium or reevaluate the risk category. This covers thus non-observable characteristics.

Even after using all available explanatory variables, it is clear that there is still heterogeneity within the risk categories or tariff cells since not we can not know the drinking habits or knowledge of the traffic rules of every driver. This can be modelled by a random effect in the statistical model. Note that in experience rating, a history with claims may also lead to more careful driving hence a lower risk of future claims, but this is beyond the scope of this thesis.

When we will develop models and first the theory of generalized linear models in the next chapter, we will need a transparent notation for these explanatory variables and their estimates. The estimate is a value given to this variable, or level of this variable, which will denote the effect on the response variable (large or small, positive or negative). This value is estimated by using the subpopulation that has this certain level of the considered variable: to estimate the difference between male and female, the population is divided in two subpopulations obviously. The number of estimates or parameters is important since each parameter that has to be estimated, shows a certain variation and the more estimates a model has, the more difficult it is to estimate accurately since the data is subdivided in
more subpopulations.

In general, such a variable will be denoted with $x$ and two subscripts $i, j$ to indicate that it is the value of this variable for policyholder $i$ (first subscript), and the value of variable $j$ (second subscript); for example $x_{i0}$ denotes the age of policyholder $i$. In case of a continuous variable, for example age, the variable can then take (in principle) values from $\mathbb{Z}$ (or even $\mathbb{R}$). If it is a categorical variable, two options of notation are possible. When coding this variable as different binary variables, there are several (second) subscripts necessary to indicate which level we are referring to. For example if there are five age classes, then we denote $x_{i1}$ for the indicator (binary variable) whether the policyholder is in age class 1 ($x_{i1} = 1$) or not ($x_{i1} = 0$), and analogously for $x_{i2}, x_{i3}, x_{i4}$ and $x_{i5}$. But one always chooses a base or reference level such that only $k - 1$ binary variables are needed for a categoric variable with $k$ levels (avoid overparameterization): obviously if $x_{i1} = x_{i2} = x_{i3} = x_{i4} = 0$ then we know that the policyholder is in age class 5. This coding method is sometimes necessary when using specific statement or statistical programs; we will not use it. A second, more direct and easier option is declaring the variable as categoric in the programming statement (see also the SAS code in Appendix B). Then simply $x_{ij}$ denotes the value of variable $j$ for observation $i$, so $x_{ij}$ is a value from a limited set $a_1, \ldots, a_k$ (for example age classes 1, 2, \ldots, 5).

When developing our models, we will use interactions: variable1*variable2, so that the estimates will depend both on the value of variable1 and variable2. This may be two continuous or categorical variables, or one of both; we just have to declare the categorical variables to make the program aware of it.

## 1.2 Bonus Malus systems

In many countries, insurers reward their good drivers and punish their bad ones by correcting the calculated premium by one factor that reflects the level of the policyholder on a bonus-malus ladder. One starts at the middle and earn points for each claim-free year, but looses (more) points when a claim (in fault) occurs. There is a maximum level of malus, where for example the premium is multiplied by a certain factor bigger than 1, and a minimum level where the multiplier is smaller than 1. So the BM is in fact a a posteriori variable that splits the risk categories again in different BM categories. In that way, BM systems can be modelled using Markov chains because of the memoryless property: the
knowledge of the past (when or how a claim occurred) does not matter for the future, only the knowledge of the present state (i.e. the BM level) - but this is beyond the scope of this thesis.

In this dataset (as in any dataset from any insurer), it is important to note however that this BM is not the true BM one should find for the policyholder. Often the insurer gives a discount in the form of a better BM, meaning that for instance young drivers which parents are insured by the same insurer, do not start at the middle but already at a lower (better) level. Or other companies reward good clients, that for example have other insurances covered by the same insurer, with points for each claim-free year in every insurance contract, that can be used to buy off the penalties from a claim, meaning that they maintain the same level of BM in stead of jumping to a worse level. This strategy is used both for attracting and bounding clients to the insurer.

The system used here is the structure of 22 steps on the ladder where one starts at 11, climbs 5 steps for each claims and descends 1 step for each claim-free year. The calculated premium for the policyholder (based on age/power of the car/ ... ) is then multiplied by a factor between \[ \frac{1}{22} \] (for BM equal to 0) and \[ 1 \] (for BM equal to 22).

Note that an accident in fault is (almost) always penalized in the same way, disregarding the severity of the resulting claim(s). This means that the insurer implicitly assumes that the number of claims and cost of a claim are independent, and that once an accident occurs, the driver’s characteristics do not influence the severity of the accident. Clearly it may be even more profitable not to report a small claim, since the increase in premium may cost more than the repair costs. This is known as the ‘hunger for bonus’ and ‘censors claim amounts and claim frequencies’ [8]. It might be a good idea to introduce levels of severity resulting in different penalties, regarding the kind and cost of damage, in the future.

Bonus-Malus systems are a tool to correct (only partly) for the phenomenon of antisymmetric information: the policyholder knows and takes advantage of information they know about their driving patterns but the insurer doesn’t. The danger for adverse selection, where only the bad drivers get an insurance, is not an issue with third party liability since this is obliged by law, but it is important for all the related (not obligatory) coverages. Moral hazard (when the insured drives less safe because he is insured) however is always a problem, especially when the safer drivers think that they are not enough rewarded for their good behaviour and have to pay (almost) the same premium as less careful drivers. A last important remark is that policyholders partly reveal information through the chosen contract: a more complete coverage will be chosen by more riskier policyholders while high
1.3 Exposure, claim frequency and severity, pure premium

A key principle used here as in actuarial ratemaking methods, is cost-based pricing, meaning that the calculated premium the insured has to pay, is actually the estimated future cost of this insurance contract for the insurer. In the pure premium approach, the estimated future costs are divided by the exposure to calculate the price of the contract (expenses are also added such as taxes, administration expenses, ...).

**Exposure** is the measure of weight one has to give to a certain value of a certain observation, because the values may otherwise be incomparable. For the frequency for example it is obvious that two contracts that each produced one claim, are not immediately comparable if one does not correct for the duration of the contract. If one contract only covers one month and the other one a whole year, there is a big difference in interpretation. Or if you compare the total cost of 1 claims or 10 claims, you need to adjust for this amount of claims to be able to compare the average cost.

So the claim frequency is the number of claims divided by the duration of the insured period measured in years, meaning that this frequency is equivalent to claims per year. The loss severity is the payment per incurred claim, so the product of the claim frequency and the loss severity is the loss per duration unit, or the loss per year, or the pure premium. The premium income or earned premium is the total of premium payments by the insured, received by the insurer; so the loss ratio is the claim amount divided by the premium income. Sometimes the term combined ratio is used where the claim amount is increased with the administrative expenses.

All these ratios are of the same type: an outcome divided by a number that measures the exposure. For the claim frequency, the exposure is the amount of time the policyholder is covered for the risk; for the claim severity, the exposure is the number of claims. The notation $Y = \frac{X}{w}$ will be used for a ratio, where we have:
1.3 Exposure, claim frequency and severity, pure premium

Table 1.1: Important ratios.

<table>
<thead>
<tr>
<th>Exposure</th>
<th>Response</th>
<th>Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Duration</td>
<td>Number of claims</td>
<td>Claim frequency</td>
</tr>
<tr>
<td>Duration</td>
<td>Claim cost</td>
<td>Pure premium</td>
</tr>
<tr>
<td>Number of claims</td>
<td>Claim cost</td>
<td>(average) Claim severity</td>
</tr>
<tr>
<td>Premium income</td>
<td>Claim cost</td>
<td>Loss ratio height</td>
</tr>
</tbody>
</table>

1.3.1 Assumptions

Several assumptions are made, we adopt them here from [5] (response is a certain ratio as in (1.1)).

1. All individual claims are independent:

   (a) Independent policies: the responses from $n$ random policies are independent. This doesn’t hold for car insurance (collisions between cars insured by the same insurer) but the effect of neglecting this should be small.

   (b) Time independence: the responses from disjoint time intervals are independent. This again is not entirely true since the winter may be more dangerous than the summer because of weather changes. But still this shouldn’t be violated too much and this assumption is necessary for substantially simplifying the model.

2. Homogeneity: the responses from policies in the same category with the same exposure have the same probability distribution.

   This is the most violated assumption: as already stressed, we try to divide the policyholder in a fair way in categories according to their risk profile, but we can’t capture all the variation since we don’t have all the needed information. Also the timing of the insurance contracts does matter (e.g. seasonal variations) within one category, but these objections apply to each category hence we assume that this violation changes the overall tariff level but not the relation between the different categories.

1.3.2 Properties

Now the consequences of this correction for exposure are explained as in [5]. Consider different policies within a tariff cell and the associated responses $X_i$ (so $i$ denotes the
1.3 Exposure, claim frequency and severity, pure premium

observation or policy). Because of the assumptions made, we can denote the mean of the response $X_i$ by $\mu$ and the variance by $\sigma^2$, both independent of $i$.

For the situation where the exposure is the number of claims $w$, so $X$ is the claim cost, we immediately have ($X = X_1 + \ldots + X_w$ so $X_j$ is the cost of claim $j$):

$$E[Y] = E\left[\frac{X}{w}\right] = \frac{1}{w}E[X] = \frac{1}{w}\sum_{i=1}^{w} E[X_i] = \frac{1}{w}wE[X_1] = \mu$$

$$Var[Y] = Var\left[\frac{X}{w}\right] = \frac{1}{w^2}Var[X] = \frac{1}{w^2}\sum_{i=1}^{w} Var[X_i] = \frac{1}{w^2}wVar[X_1] = \frac{\sigma^2}{w}.$$ 

For the situation where the exposure is the duration or premium, the same results are valid when $\mu$ and $\sigma^2$ denote then the expected value and variance for a response with exposure 1. To see this we suppose that the total exposure $w$ is a rational number $m/n$ so consists of $m$ time intervals of equal length or equal premium income $1/n$. The responses in these intervals, $X_1, \ldots, X_m$, are thus independent and identically distributed variables with exposure $w_i = 1/n$. So by adding $n$ such responses, we get a variable $Z$ with exposure 1 and by assumption, $E[Z] = \mu$ and $Var[Z] = \sigma^2$. Now

$$E[X_j] = E[X_1] = \frac{1}{n}E[Z] = \frac{1}{n}\mu,$$

$$Var[X_j] = Var[X_1] = \frac{1}{n}Var[Z] = \frac{1}{n}\sigma^2.$$ 

Then we have with $X = X_1 + \ldots + X_m$:

$$E[Y] = E\left[\frac{X}{w}\right] = \frac{1}{w}E[X] = \frac{1}{w}\sum_{i=1}^{m} E[X_j] = \frac{1}{w}mE[X_1] = \frac{1}{w}m\frac{\mu}{n} = \mu$$

$$Var[Y] = Var\left[\frac{X}{w}\right] = \frac{1}{w^2}Var[X] = \frac{1}{w^2}\sum_{i=1}^{m} Var[X_j] = \frac{1}{w^2}mVar[X_1] = \frac{1}{w^2}m\frac{\sigma^2}{n} = \frac{\sigma^2}{w}.$$ 

The transition to all real $w$ results from taking the limit for a sequence of rational numbers that converges to $w$ (it is well known that for every real number such a sequence exists).

The important consequence is thus that we should always use weighted variances when modelling a ratio.

The problem with liability insurance is that the settlement of larger claims often takes several years (for example if someone is hurt, it may take years before the consequences of the injuries are fully understood and can be translated into a cost). In this case, loss
development factors can be used to estimate these costs. In the dataset here the final costs are given per policy; so when a policy produced two claims we only know the total cost of these two claims together.

1.4 Claim counts

1.4.1 Poisson distribution

The binomial distribution is probably the most known discrete distribution: it describes the number of successes (mostly denoted by 1 and failure by 0) when performing an experiment $n$ times, where the chance of succes in each experiment is $p$. We will always denote with $p(y)$ the probability of outcome $y$.

$$Y \overset{d}{=} \text{Binom}(n, p) \iff p(y) = \binom{n}{y} y^p (1-p)^{n-y} \quad (y = 0, 1, \ldots, n).$$

If $n$ is large enough and $p$ is not too close to 0 or 1 (meaning that the skewness is not too great), then the normal distribution is a good approximation. If $n$ is large enough and $p$ is small, which is clearly the case for many insurance coverages, then the Poisson distribution is a good approximation. This distribution is characterized by one parameter $\lambda$ which is the distribution’s mean and variance (in the approximation of the binomial case, $\lambda = np$).

$$Y \overset{d}{=} \text{Pois}(\lambda) \iff p(y) = \exp(-\lambda) \frac{\lambda^y}{y!}.$$ (1.1)

This can be seen when we denote $N \overset{d}{=} \text{Binom}(n, \frac{\lambda}{n})$ and take the limit $n \to +\infty$:

$$p(0) = \left(1 - \frac{\lambda}{n}\right)^n \to \exp(-\lambda)$$

$$\frac{p(k+1)}{p(k)} = \frac{\frac{n-k}{n} \frac{\lambda}{n}}{1 - \frac{\lambda}{n}} \to \frac{\lambda}{k+1}.$$  

Using the identity $\sum_{k=0}^{+\infty} \frac{\lambda^k}{k!} = \exp(\lambda)$, one can easily compute that

$$Y \overset{d}{=} \text{Pois}(\lambda) \Rightarrow E[Y] = \lambda, \ E[Y^2] = \lambda + \lambda^2, \ Var[Y] = E[Y^2] - E^2[Y] = \lambda.$$  

The skewness then is $\gamma_Y = 1/\sqrt{\lambda}$ so as $\lambda$ increases, the distribution gets less skewed (nearly symmetric for $\lambda = 15$). The probability generating function of the Poisson distribution has a very simple form:

$$\phi_Y(z) = E\left[z^Y\right] = \sum_{k=0}^{+\infty} \exp(-\lambda) \frac{(\lambda z)^k}{k!} = \exp(\lambda(z - 1)).$$
Since the pgf of the sum of two distributions is the product of the pgf’s, the sum of two independent Poisson distributions \( Y_1 \overset{d}{=} \text{Pois}(\lambda_1) \) and \( Y_2 \overset{d}{=} \text{Pois}(\lambda_2) \) is again Poisson distributed, with parameter the sum of the Poisson parameters \( Y_1 + Y_2 \overset{d}{=} \text{Pois}(\lambda_1 + \lambda_2) \).

### 1.4.2 Mixed Poisson distribution

In TPL the underlying population to describe is not homogeneous, and unobserved heterogeneity results in excess zeros and (almost always observed) heavy upper tails. A mixed Poisson model may be more appropriate, where a random variable is introduced in the mean: conditionally on this random variable, the distribution is Poisson. Mixture models can combine only different discrete or only continuous distributions, as well as discrete and continuous distributions - this is typically the case when a population is heterogeneous and consists of subpopulations whose distribution can be simplified. A mixture of Poisson distributions means that these subpopulations \( i \) are Poisson distributed with a certain parameter \( \lambda_i \), and one doesn’t know for a fact to which subpopulation an observation belongs, but does know the probability \( p_i \) that it comes from the \( i \)th subpopulation. If we now denote with \( \Theta \) the unobservable random variable such that the mean frequency is multiplied by this effect, then given \( \Theta = \theta \), the probability is Poisson distributed [8]:

\[
p(Y = k|\Theta = \theta) = p(k|\lambda \theta) = \exp(-\lambda \theta) \frac{(\lambda \theta)^k}{k!}.
\]

In general, \( \Theta \) is not discrete or continuous but of mixed type and by definition of the expectance and distribution function \( F_\Theta \) of \( \Theta \) there holds:

\[
p(Y = k) = E[p(k|\lambda \Theta)] = \int_0^\infty \exp(-\lambda \theta) \frac{(\lambda \theta)^k}{k!} dF_\Theta(\theta).
\]

The notation for this distribution is \( Y \overset{d}{=} \text{MPois}(\lambda, \Theta) \). The condition \( E[\Theta] = 1 \) ensures that \( E[Y] = \lambda \):

\[
E[Y] = \sum_{k=0}^{+\infty} kp(k) = \sum_{k=0}^{+\infty} k \int_0^\infty \exp(-\lambda \theta) \frac{(\lambda \theta)^k}{k!} dF_\Theta(\theta)
\]

\[
= \int_0^\infty \lambda \theta \sum_{k=1}^{+\infty} \exp(-\lambda \theta) \frac{(\lambda \theta)^{(k-1)}}{(k-1)!} dF_\Theta(\theta)
\]

\[
= \int_0^\infty \lambda \theta \exp(-\lambda \theta) \exp(\lambda \theta) dF_\Theta(\theta)
\]

\[
= \lambda E[\Theta] = \lambda.
\]

Or more briefly:

\[
E[N] = E[ E[N|\Theta] ] = E[\lambda \Theta] = \lambda.
\]
1.4 Claim counts

Properties

If \( Y \overset{d}{=} MPois(\lambda, \Theta) \) then its variance exceeds its mean - mixed Poisson distributions are thus overdispersed:

\[
E [Y^2] = E \left[ E [Y^2 \| \Theta] \right] \\
= \int_0^{+\infty} \left( Var [Y \| \Theta] + E^2 [Y \| \Theta] \right) dF_\Theta(\theta) \\
= \int_0^{+\infty} (\lambda \Theta + \lambda^2 \Theta^2) dF_\Theta(\theta) \\
= \lambda E[\Theta] + \lambda^2 E[\Theta^2]
\]

thus

\[
= \lambda E[\Theta] + \lambda^2 E[\Theta^2] - \lambda^2 E[\Theta] \\
= \lambda + \lambda^2 Var[\Theta] \geq \lambda = E[Y].
\]

Also the probability of observing a zero is bigger than observing one in the Poisson distribution with the same mean \( \lambda \). This can be proven with Jensen’s inequality \( E[\phi(X)] \geq \phi(E[X]) \) for any random variable \( X \) and convex function \( \phi \).

\[
Pr(Y = 0) = \int_0^{+\infty} \exp(-\lambda \theta) dF_\Theta(\theta) \geq \exp \left( - \int_0^{+\infty} \lambda \theta dF_\Theta(\theta) \right) = \exp(-\lambda).
\]

Moreover the mixed Poisson distribution has a thicker right tail than the Poisson distribution with the same mean. Shaked (1980) proved that if \( Y \overset{d}{=} MPois(\lambda, \Theta) \) there exist two integers \( 0 \leq k_0 < k_1 \) such that

\[
Pr(Y = k) \geq \exp(-\lambda) \frac{\lambda^k}{k!}, \quad k = 0, 1, \ldots, k_0,
\]

\[
Pr(Y = k) \leq \exp(-\lambda) \frac{\lambda^k}{k!}, \quad k = k_0 + 1, \ldots, k_1,
\]

\[
Pr(Y = k) \geq \exp(-\lambda) \frac{\lambda^k}{k!}, \quad k \geq k_1 + 1.
\]

The pgf of \( Y \overset{d}{=} MPois(\lambda, \Theta) \) can be expressed with the moment generating function
1.5 Claim severity

$M_\Theta(z) = E \left[ \exp(z\Theta) \right]$

$\phi_Y(z) = E \left[ z^Y \right] = \sum_{k=0}^{+\infty} p(Y = k) z^k$

$= \int_0^\infty \exp(-\lambda \theta) \sum_{k=0}^{+\infty} \frac{(z \lambda \theta)^k}{k!} dF_\Theta(\theta)$

$= \int_0^\infty \exp(-\lambda \theta) \exp(z \lambda \theta) dF_\Theta(\theta)$

$= E \left[ \exp(\lambda(z - 1)\Theta) \right] = M_\Theta(\lambda(z - 1))$.

From this identity, it is also clear that the mixed Poisson distribution is known if and only if $F_\Theta$ is known and that two mixed Poisson distributions with the same parameter $\lambda$ are the same if and only if the $F_\Theta$’s are the same.

1.4.3 Compound Poisson distribution

In general, a compound Poisson distribution is the sum of a number of random variables, which are independent and identically distributed, where the number is Poisson distributed. This can be used to model pure premium where it represents then a sum of claims, which are for example Gamma distributed (see also Tweedie models in the next chapter). Moreover, this compound Poisson distribution is then a mixed distribution since the probability at zero is positive (i.e. the probability that the number of random variables are all zero or that the number itself is zero) and the distribution on the real numbers is continuous. So it results actually from the combination of a discrete distribution (probability of being zero or not) and a continuous distribution.

1.5 Claim severity

Once the number of claims is estimated, the claim severity can be modelled: the claim frequency is analyzed, conditionally on the number of claims (which is exactly the exposure). The distribution of claim severity should be positive and right-skewed; the gamma distribution $G(\alpha, \beta)$ has become quite standard here. This implies that the coefficient of variation $Var^{1/2}/E$ is constant. So the density function of one claim (the subscript $i$ stresses the fact that this depends of the characteristics of the observation itself, the subscript $j$ denotes the defined distribution) is given by

$f_j(y_i) = \frac{\beta_j^{\alpha_j}}{\Gamma(\alpha_j)} y_i^{\alpha_j-1} e^{-\beta_j y_i}$
where \( \alpha_j > 0 \) is the index or shape parameter and \( \beta_j > 0 \) the scale parameter. Using the identity
\[
\int_0^{+\infty} x^s e^{-tx} dx = \frac{\Gamma(s + 1)}{t^{s+1}}
\]
one easily finds that \( E[G(\alpha, \beta)] = \frac{\alpha}{\beta} \) and \( \text{Var}[G(\alpha, \beta)] = \frac{\alpha}{\beta^2} = E[G]/\beta \). So the coefficient of variation is \( 1/\sqrt{\alpha} \) and depends only on the index parameter \( \alpha \).

The mgf of \( G(\alpha, \beta) \) is given by
\[
m_G(t) = E[e^{tG}] = \left( \frac{\beta}{\beta - t} \right)^{\alpha} \quad (t < \beta).
\]

Since the mgf of a sum of independent random variables is the product of the mgf’s, it is immediately clear that the sum of independent gamma distributions with the same scale parameter \( \beta \) is again gamma distributed where the new index parameter is the sum of all index parameters. So if \( X_i = \sum_{j=1}^{n} X_{ij} \) (think of multiple claims of policyholder \( i \)) where \( X_{ij} \overset{d}{=} G(\alpha_i, \beta_i) \), then \( X_i \overset{d}{=} G(n\alpha_i, \beta_i) \). Hence the density of \( Y_i = X_i/n \) (think of the average claim) is then
\[
f_{Y_i}(y) = nf_{X_i}(n) = \frac{\beta_{i}^{n\alpha_i}}{\Gamma(n\alpha_i)}(ny)^{n\alpha_i-1}e^{-\beta_i ny}
\]
\[
= \frac{(n\beta_i)^{n\alpha_i}}{\Gamma(n\alpha_i)}y^{n\alpha_i-1}e^{-\beta_i ny}
\]
so that \( Y_i \overset{d}{=} G(n\alpha_i, n\beta_i) \) with the same expectation as \( X_{ij} \), namely \( \alpha_i/\beta_i \).
Chapter 2

Generalized Linear Models

We can distinguish two main approaches in car insurance: one where the observable covariates are disregarded and all the individual characteristics are assumed to be represented by random variables, the other one tries to explain the variation without random effects hence only by the observable differences. For example, when estimating the Poisson parameter as in the previous section for the whole population or the subpopulation, the first approach is used. It is mostly interesting to combine both views.

Regression models try to capture the relation between the response variable (the variable one is trying to predict, for example the claim frequency) and the explanatory variables (or predictors or covariates). This relation is expressed in a distribution function which produces predicted values for the response variable and the parameters of this distribution function are obtained by optimizing a measure of fit. It is of course crucial to use appropriate covariates that capture the variation and different categories the best. As already mentioned, there is still unexplained variation between different categories hence random effects can be added to the predictors which indeed combines then the two approaches.

All analyses will be made with the help of and based on GLM. Nelder and Wedderburn discovered that regression models where the response variable is distributed as a member of the exponential family share the same characteristics. In contrary to the classical normal linear regression, there are less restrictions here: in addition to the wide gamma of possible response distributions, the variance need not to be constant (heteroscedasticity is allowed) and the relation between the predicted values (or fitted values) and the predictors need not to be linear. We now describe all this in detail.
2.1 General model

The exponential dispersion family contains all distributions whose frequency function is of the form

\[
f_{Y_i}(y_i; \theta_i, \phi) = \exp \left[ \frac{y_i \theta_i - b(\theta_i)}{\phi/w_i} + c(y_i, \phi, w_i) \right]. \tag{2.1}
\]

Here \(y_i\) is the observed response of a certain observation with certain characteristics, the natural or canonical parameter \(\theta_i\) is allowed to vary with these characteristics while the dispersion parameter \(\phi > 0\) does not and \(w_i \geq 0\) is the weight associated to this observation.

The parameter \(\theta_i\) takes values in an open set (f.ex. \(0 < \theta < 1\)) and the function \(b(\theta_i)\) is the cumulant function and is assumed twice continuously differentiable, with invertible second derivative because of the following properties:

\[
E[Y_i] = \mu_i = b'(\theta_i) \tag{2.2}
\]

\[
\text{Var}[Y_i] = \frac{\phi}{w_i} b''(\theta_i) = \frac{\phi}{w_i} V(\mu_i) \text{ with } V \text{ the variance function.}
\]

This can be proven using the loglikelihood

\[
L = \ln f_{Y_i}(y_i; \theta_i, \phi) = w_i \frac{y_i \theta_i - b(\theta_i)}{\phi} + c(y_i, \phi, w_i) \tag{2.3}
\]

since there holds under the regularity conditions that

\[
0 = E \left[ \frac{d}{d\theta_i} L \right] = E \left[ \frac{w_i}{\phi} [y_i - b'(\theta_i)] \right] \Rightarrow \mu_i = E[Y_i] = b'(\theta_i),
\]

\[
E \left[ \left( \frac{d}{d\theta_i} L \right)^2 \right] = -E \left[ \frac{d^2}{d\theta_i^2} L \right] \Rightarrow E \left[ \frac{w_i^2}{\phi^2} (y_i - b'(\theta_i))^2 \right] = -E \left[ \frac{b''(\theta_i)}{\phi} w_i \right]
\]

\[
\Rightarrow \text{Var}[Y_i] = \frac{\phi}{w_i} b''(\theta_i).
\]

Another proof, which we will elaborate here now, is perhaps more natural and given in \[5\]. This uses the cumulant generating function \(\Psi\), which is the logarithm of the moment generating function \(M_Y(t) = E[\exp(tY)]\) (if this expectation is finite at least for \(t \in \mathbb{R}\) in a neighborhood of zero). In case of the exponential family (we drop the subscript \(i\) here):

\[
\Psi_Y(t) = \ln M_Y(t) = \ln E[\exp(tY)] = \ln \int \exp(ty) f_Y(y; \theta, \phi) dy.
\]

For continuous distributions we find that

\[
\int \exp(ty) f_Y(y; \theta, \phi) dy = \int \exp \left[ \frac{y(\theta + t\phi/w) - b(\theta + t\phi/w)}{\phi/w} + c(y, \phi, w) \right] dy
\]

\[
= \exp \left[ \frac{b(\theta + t\phi/w) - b(\theta)}{\phi/w} \right] \int \exp \left[ \frac{y(\theta + t\phi/w) - b(\theta + t\phi/w)}{\phi/w} + c(y, \phi, w) \right] dy.
\]
2.1 General model

Now this last integral equals one if \( \theta + t\phi/w \) is in the parameter space, so at least for \( t \) in a neighborhood of zero. Note that the same result is obtained for the discrete case where the integrals are then changed in sums. So the cgf exists for any member of the exponential family, at least for \( |t| < \delta \) for some \( \delta > 0 \), and is given by

\[
\Psi_Y(t) = \frac{b(\theta + t\phi/w) - b(\theta)}{\phi/w}.
\]

This also shows where the function \( b(\theta) \) got his name as cumulant function. The so called cumulants are obtained by differentiating and setting \( t = 0 \): the first derivative gives the expected value, the second gives the variance (recall that \( b \) is assumed twice differentiable). We derive

\[
\Psi'(t) = b'(\theta + t\phi/w) \\
\implies E[Y] = \Psi'(0) = b'(\theta) \\
\Psi''(t) = b''(\theta + t\phi/w)\phi/w \\
\implies \text{Var}[Y] = \Psi''(0) = b''(\theta)\phi/w.
\]

And also since \( b' \) is assumed invertable, there holds that \( \theta = (b')^{-1}(\mu) \) so that

\[
\text{Var}[Y] = b''(b'^{-1}(\mu))\phi/w = V(\mu)\phi/w
\]

with \( V(\mu) \) the variance function.

If \( b, \theta_i \) and \( \phi \) are specified, the distribution is completely determined (\( c \) is not important for GLM theory). This family contains the normal, binomial, Poisson, gamma, inverse Gaussian, ... distribution: \( f_Y \) is the probability density function in the continuous case and the probability mass function in the discrete case. Note that for fixed \( \phi \), this family is the so called one-parameter exponential family. The lognormal distribution and Pareto distribution are examples of distribution that don’t belong to the exponential dispersion family.

Recalling the notation \( x_{ij} \) for explanatory variables of observation \( i \) with levels \( j \), we can define the score function or linear predictor of an observation:

\[
\text{score}_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}
\]

where \( \beta_j \) are the unknown regression coefficients, to be estimated from the data. The coefficient \( \beta_j \) indicates how much weight is given to the \( j \)th covariate, \( \beta_0 \) is the intercept.
This score function is related to the mean of the distribution function by the following relation:

\[ \text{score}_i = g(\mu_i) \]

where \( g \) is the link function. So the linear (or additive, \( g = 1 \)) and multiplicative model are special cases. The link function \( g \) is called the canonical link if it satisfies

\[ \theta_i \equiv g(\mu_i) = g(E[Y_i]) = \text{score}_i. \]

These are used the most because they guarantee maximal information, simplify estimating and offer a simpler interpretation for the regression parameters.

### 2.2 Estimators

When we have a sample of observations \( y_1, y_2, \ldots, y_n \), the estimators for \( \beta_1, \ldots, \beta_p \), denoted with \( \hat{\beta}_j \) (1 \( \leq j \leq p \)), are solutions of the \( p \) equations (recall the loglikelihood (2.3))

\[
\frac{d}{d\beta_j} \mathcal{L} = \sum_{i=1}^{n} \frac{d}{d\beta_j} \mathcal{L}_i = \sum_{i=1}^{n} \frac{d}{d\beta_j} \left( w_i \frac{y_i \theta_i - b(\theta_i)}{\phi} + c(y_i, \phi, w_i) \right)
\]

which can be further elaborated (score \( i = \eta_i \) for simplicity of notation, \( \delta_{ij} = 1 \) if \( i = j \) and zero otherwise):

\[
\sum_{i=1}^{n} \frac{d}{d\beta_j} \mathcal{L}_i = \sum_{i=1}^{n} \sum_{k} \sum_{l} \sum_{m} \frac{d\mathcal{L}_i}{d\theta_k} \frac{d\mu_l}{d\eta_m} \frac{d\eta_m}{d\beta_j}
\]

\[
= \sum_{i=1}^{n} \sum_{l} \sum_{m} \left( \frac{w_i y_i - b'(\theta_i)}{\phi} \right) \delta_{ik} \frac{d\theta_l}{d\mu_l} \frac{d\mu_l}{d\eta_m} \frac{d\eta_m}{d\beta_j}
\]

\[
= \sum_{i=1}^{n} \sum_{l} \sum_{m} \frac{w_i y_i - b'(\theta_i)}{\phi} \left( \frac{d\mu_l}{d\theta_l} \right)^{-1} \frac{d\mu_l}{d\eta_m} \frac{d\eta_m}{d\beta_j}
\]

\[
= \sum_{i=1}^{n} \sum_{m} \frac{w_i y_i - \mu_i}{\phi} \frac{1}{b''(\theta_i)} \delta_{ij} \frac{d\mu_l}{d\eta_m} \frac{d\eta_m}{d\beta_j}
\]

\[
= \sum_{i=1}^{n} \sum_{l} \frac{w_i y_i - \mu_i}{\phi} \frac{1}{V(\mu_i)} \frac{1}{g'(\mu_i)} \frac{d\eta_m}{d\beta_j}
\]

\[
= \sum_{i=1}^{n} \frac{w_i y_i - \mu_i}{\phi} \frac{1}{V(\mu_i)} \frac{x_{ij}}{g'(\mu_i)}.
\]

So we can multiply with \( \phi \) and then the estimators \( \hat{\beta}_j \) are solutions of the maximum likelihood equations

\[
\sum_{i=1}^{n} \frac{w_i y_i - \mu_i}{V(\mu_i)} \frac{x_{ij}}{g'(\mu_i)} = 0 \quad (1 \leq j \leq p).
\]
Don’t forget that at the same time score$_i = g(\mu_i)$ has to be fulfilled! So these equations are nonlinear with respect to $\beta_j$ hence iterative methods have to be used to get numerical solutions.

### 2.3 Confidence interval

First we introduce the *Fisher information matrix* $\mathcal{I}$ of a set of estimators $\beta_j$ ($1 \leq j \leq p$). Its elements are defined as

$$
\mathcal{I}_{jk} = -E \left[ \frac{d^2}{d\beta_j d\beta_k} \mathcal{L} \right] = -E [\mathcal{H}_{jk}]
$$

where $\mathcal{H}$ is called the Hessian matrix. Recalling previous calculations, we thus have

\[
\frac{d^2}{d\beta_j d\beta_k} \mathcal{L} = \frac{d}{d\beta_k} \left( \sum_i w_i \frac{y_i - \mu_i}{\phi} \frac{1}{V(\mu_i) g'(\mu_i)} x_{ij} \right) \\
= \sum_i \frac{w_i}{\phi} \frac{d}{d\mu_i} \left( \frac{y_i - \mu_i}{V(\mu_i) g'(\mu_i)} \right) x_{ij} \frac{d\mu_i}{d\eta} \frac{d\eta}{d\beta_k} \\
= \sum_i \frac{w_i}{\phi} \frac{d}{d\mu_i} \left( \frac{y_i - \mu_i}{V(\mu_i) g'(\mu_i)} \right) x_{ij} \frac{1}{g'(\mu_i)} x_{ik} \\
= \sum_i \frac{w_i}{\phi} \frac{-V(\mu_i) g'(\mu_i) - (y_i - \mu_i)(V'(\mu_i) g'(\mu_i) + V(\mu_i) g''(\mu_i))}{(V(\mu_i) g'(\mu_i))^2} x_{ij} \frac{1}{g'(\mu_i)} x_{ik} \\
= \sum_i \frac{w_i}{\phi} \frac{-V(\mu_i) g'(\mu_i) - (y_i - \mu_i)(V'(\mu_i) g'(\mu_i) + V(\mu_i) g''(\mu_i))}{(V(\mu_i))^2 (g'(\mu_i))^3} x_{ij} x_{ik}.
\]

When taking the expectation, the second term disappears since $E(y_i) = \mu_i$ so that

$$
\mathcal{I}_{jk} = \sum_i \frac{w_i}{\phi} \frac{1}{V(\mu_i) (g'(\mu_i))^2} x_{ij} x_{ik}.
$$

So the information grows linearly in $w_i$, and is inverse proportional to $\phi$.

From general ML estimation theory, the MLE’s are under general conditions, asymptotically normally distributed and unbiased, with covariance matrix equal to the inverse of the Fisher information matrix $\mathcal{I}$. So the resulting approximation (in distribution) is

$$
\hat{\beta} \approx N(\beta; \mathcal{I}^{-1}).
$$

So a confidence interval for the estimated $\beta$ can be computed: if $b_{ij}$ denotes the matrix element $(\mathcal{I}^{-1})_{ij}$, then the $(1 - \alpha)$% confidence interval for $\beta_j$ is given by

\[
\left[ \hat{\beta}_j - z_{1-\alpha/2} \sqrt{b_{jj}}, \hat{\beta}_j + z_{1-\alpha/2} \sqrt{b_{jj}} \right]
\] (2.6)
2.4 Estimation of $\phi$

where $z_\alpha$ is the $\alpha$ quantile of the standard normal distribution. Herein $I$ needs to be estimated as well of course, by inserting the estimates $\hat{\mu}_i$ and $\hat{\phi}$. Confidence intervals are very important since they indicate the precision of the estimates: the smaller the interval is, the more reliable the estimator is.

In chapter four, we will estimate the pure premium (ultimate cost/duration) by multiplying the estimates for the claim frequency and claim severity, how will we obtain the confidence interval for this pure premium estimate then? We adopt here the approach from [5]. Denote the variance of the estimator of the claim frequency $\hat{\beta}^F$ by $\text{Var} [\hat{\beta}^F]$ (so this is the estimator of $\ln(\mu)$, not yet the relativities $\exp(\hat{\beta}^F)$) and $\text{Var} [\hat{\beta}^S]$ for the variance of the estimator of the claim severity $\hat{\beta}^S$. Then we want to determine $\text{Var} [\hat{\beta}^P]$, the variance of the estimator of the pure premium $\hat{\beta}^P$. Because the severity depends on the number of claims, there could be some dependence between the claim frequency and severity. Furthermore the analysis of the severity is made conditionally on the number of claims. As already noted, the estimates $\hat{\beta}^X$ here are approximately unbiased: $E [\hat{\beta}^S|\text{nclaim}] \approx \beta^S$. Then there also holds that $\text{Var} [\hat{\beta}^F|\text{nclaim}] = 0$ because $\hat{\beta}^F$ is only based on nclaim, so one concludes:

$$\text{Var} [\hat{\beta}^P] = \text{Var} \left[ E [\hat{\beta}^F + \hat{\beta}^S|\text{nclaim}] \right] + E \left[ \text{Var} [\hat{\beta}^F + \hat{\beta}^S|\text{nclaim}] \right]$$

$$\approx \text{Var} [\hat{\beta}^F] + E \left[ \text{Var} [\hat{\beta}^S|\text{nclaim}] \right]$$

$$\approx \text{Var} [\hat{\beta}^F] + \text{Var} [\hat{\beta}^S|\text{nclaim}],$$

where the conditional variance $\text{Var} [\hat{\beta}^S|\text{nclaim}]$ is actually the variance one gets in the GLM analysis of claim severity. So an estimate of $\text{Var} [\hat{\beta}^P]$ turns out to be the sum of the variances

$$\hat{\text{Var}} [\hat{\beta}^P] = \hat{\text{Var}} [\hat{\beta}^F] + \hat{\text{Var}} [\hat{\beta}^S|\text{nclaim}].$$

(2.7)

With this variance, or the standard error $\sqrt{\text{Var}}$, one can compute the CI by the formula (2.6).

2.4 Estimation of $\phi$

As seen in (2.2), the parameter $\phi$ scales the relationship between the variance and the mean. In practice $\phi$ is often unknown and needs to be estimated in order to be able to
compute the Fisher information matrix hence the confidence intervals. Several options are possible, but in the literature it seems that the estimator using the Pearson’s statistic is mostly recommended and apparently more robust against model error \cite{5}, \cite{12}, \cite{10}. SAS uses by default the ML estimator, but by using a certain option \textit{pscale}, one can get estimations with the Pearson’s statistic.

The Pearson’s chi-square statistic $X^2$ is a classic measure of the goodness of fit of a statistical model:

$$X^2 = \frac{1}{\phi} \sum_i w_i \frac{(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)}.$$  

Now it follows from statistical theory that $X^2$ is approximately $\chi^2_{n-r}$ distributed, with $r$ the number of estimated parameters ($\beta$’s). So $E(X^2) \approx n - r$ and an approximately unbiased estimator of $\phi$ is hence

$$\hat{\phi}_X = \frac{\phi X^2}{n - r} = \frac{1}{n - r} \sum_i w_i \frac{(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)}.$$  

Notice that since $I_{\phi}^{-1} \propto \phi$, the smaller the confidence intervals become as $\phi$ becomes smaller (for example for larger $n$ or smaller $r$).

### 2.5 Poisson model

For the Poisson distribution we have, for $Y_i = X_i/w_i$ with $X_i \overset{d}{=} \text{Pois}(\lambda_i w_i)$ (since $\lambda_i$ is the expectation if $w_i = 1$) from \eqref{1.1}:

$$p(y_i) = f_{Y_i}(y_i) = P(X_i = w_i y_i) = \exp(-\lambda_i w_i) \frac{(\lambda_i w_i)^{w_i y_i}}{(w_i y_i)!}$$

$$= \exp(-\lambda_i w_i) \exp \left[ w_i y_i \ln(\lambda_i w_i) - \ln((w_i y_i)!) \right]$$

$$= \exp \left[ w_i \left( y_i \ln(\lambda_i) - \lambda_i \right) + \left[ w_i y_i \ln(w_i) - \ln((w_i y_i)!) \right] \right].$$  

So from \eqref{2.1} it is clear that $\phi = 1$, $\theta_i = \ln(\lambda_i)$, $b(\theta_i) = \exp(\theta_i)$ and the parameter space is open: $\lambda_i > 0$ or $\theta_i \in \mathbb{R}$. And indeed the expressions in \eqref{2.2} can be verified for the mean $\mu_i = b'(\theta_i) = \exp(\ln \lambda_i) = \lambda_i$ and the variance function $V(\mu_i) = b''(\theta_i) = \exp(\ln \lambda_i) = \lambda_i = \mu_i$ ($V \equiv 1$).

The canonical link function for the Poisson distribution is $g = \ln$, so indeed the positive claim frequency is transformed in a score function that can have values in $\mathbb{R}$. In case the
2.5 Poisson model

response is Poisson distributed, we thus have

\[ \exp(\text{score}_i) = \exp(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}) = \lambda_i = E[Y_i] \]  \hspace{1cm} (2.10)

for the annual expected claim frequency for policyholder \( i \). Hence the ML equations (2.4) reduce to

\[ \sum_{i=1}^{n} \frac{d}{d\beta_j} \mathcal{L}_i = \sum_{i=1}^{n} x_{ij} w_i (y_i - \lambda_i) = 0 \hspace{0.5cm} (1 \leq j \leq p) \]  \hspace{1cm} (2.11)

For example in this case, the number of claims produced by a certain policyholder \( X \) can be predicted as the (random) outcome of a Poisson distribution where its parameters is estimated by the age of the car and the age of the policyholder, so if these ages are filled in for \( X \), the parameter can be calculated hence the distribution function is known.

From (2.10) it is also clear that the resulting tariff is a multiplicative tariff: the reference class, for which all variables equal zero, have \( \lambda_i = \exp(\beta_0) \), then for each non-zero (continuous or categorical) covariate \( x_{ij} \) this becomes:

\[ \lambda_i = \exp(\beta_0) \prod_{j|x_{ij}\neq 0} \exp(\beta_j x_{ij}) = \exp(\beta_0 + \sum_{j|x_{ij}\neq 0} \beta_j x_{ij}) \]  \hspace{1cm} (2.12)

so the impact of the \( j \)th covariate on the annual claim frequency is not \( \beta_j x_{ij} \), but the factor \( \exp(\beta_j x_{ij}) \). Hence if \( \beta_j > 0 \) (assuming \( x_{ij} \geq 0 \)), this factor increases the frequency (factor is bigger than 1), if \( \beta_j < 0 \) the frequency decreases.

Note that we can merge two or more tariff cells with the same expectation into one since each cell has a relative Poisson distribution which is reproductive. Suppose \( Y_i \) is the claim frequency in the cell with exposure \( w_i \) \((i = 1, 2)\). Then the claim frequency in the new (merged) cell will be

\[ Y = \frac{w_1 Y_1 + w_2 Y_2}{w_1 + w_2} \]

where the nominator is a linear combination of Poisson distributions so again Poisson distributed and the expectation is clearly \( \lambda = E[Y_1] = E[Y_2] \). For the discussed ratios in (1.1) it is natural that their distribution is closed under this kind of averaging; it turns out that all distributions from the exponential dispersion family are reproductive. So the weighted average of independent random variables with the same function \( b \), mean and \( \phi \), belongs to the same distribution with the same \( b \), mean and \( \phi \). This is of course very useful in the context of a tarification: for Poisson distributed cells, those with the same expectation, thus the same tariff, may be merged into one cell. This results in the same Poisson distribution with the common mean as parameter but reduces the number of cells or tariff levels!
Over- and underdispersion

The dispersion parameter is in theory equal to 1 for the Poisson distribution, so the estimator \( \hat{\phi} = \frac{\phi X^2}{n - r} = \frac{1}{n - r} \sum_i w_i \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i} \) should be around 1. In reality, this is often larger because of unobserved heterogeneity: the data displays more variation than theoretically assumed. Important explanatory variables may not yet have been measured or included in the model and this is often taken into account by introducing a random effect in the model (credibility construction).

Because of the relation \( \text{Var} \propto \phi \) for the exponential family (2.2) one speaks of overdispersion if \( \hat{\phi} > 1 \), and underdispersion if \( \hat{\phi} < 1 \).

2.6 Gamma distribution

To more easily prove that the gamma distribution is a member of the exponential dispersion family, a reparametrization takes place [5]. Defining \( \mu_i = \alpha_i / \beta_i \) and \( \phi_i = 1 / \alpha_i \), the density (1.2) becomes

\[
 f_{Y_i}(y) = f_{Y_i}(y; \mu_i, \phi_i) = \frac{1}{\Gamma(n/\phi_i)} \left( \frac{n}{\mu_i \phi_i} \right)^{n/\phi_i} y^{(n/\phi_i)-1} e^{-ny/\mu_i \phi_i} 
\]

(2.13)

From this it is clear that the gamma distribution fulfills (2.1) with \( w_i = n, \theta_i = -1/\mu_i, b(x) = -\ln(-x) \). Easily one can check that the variance is divided by the weights: \( \text{Var} [Y_i] = n\alpha / (n\beta)^2 = \phi_i \mu_i^2 / n \). Using the link function \( g = \ln \), the ML equations (2.4) become

\[
 \sum_{i=1}^{n} w_i \frac{y_i - \mu_i}{\mu_i^2} x_{ij} = \frac{1}{\mu_i} \sum_{i=1}^{n} w_i \frac{y_i - \mu_i}{\mu_i} x_{ij} = 0 \ (1 \leq j \leq p). 
\]

(2.14)

2.7 Multiplicative tariff

As explained for the Poisson example, we will always assume the goal is a multiplicative model, meaning that the expected outcome is multiplied with a certain factor to know the
effect of a predictor, instead of increasing or decreasing the expected outcome with a certain value (additive model). As a consequence we are not that interested in the $\beta_i$, but in the relativities: the (multiplicative) factors, $\exp(\beta_j)$ as in (2.12). Moreover predictors don’t interact with each other: if the expected claim frequency is twice as much for young drivers (versus older drivers) then this is true for every level of every other predictor (regardless the sex for example). If this seems not true for certain levels of another predictor, one should consider the interaction of the concerning predictors: then all possible combinations of the levels of the predictors get another parameter. For example, it is well known that young drivers in general produce more claims, but young male drivers have a much higher accident rate than young female drivers, so the size of the effect should be separated for gender since we will otherwise underestimate the claim probability for males and overestimate it for females. Instead of $\beta_{sex}$ and $\beta_{age}$ you then use $\beta_{female, age}$ and $\beta_{male, age}$ (simplified notation).

But this will be elaborately discussed in chapter five, when searching for interactions in the dataset.

Note also that when we will determine the premium of the different categories, such a multiplicative model is very transparent since the intercept $\beta_0$ can be changed according to the desired base premium without changing the relativities between the different risk cells. After determining the parameters $\beta_i$ for the chosen predictors, $\beta_0$ can be increased or decreased such that over all categories together, the required total premium income is achieved.

### 2.8 Offset

A first correction that can be made for data with different time periods, is an offset. This is a continuous variable - one that is not used as an explanatory variable - that is summed up with the score function and marks a different starting value for the prediction. Then we have

$$g(\mu_i) = \text{score}_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \text{offset}_i$$

Note that can also be used when estimation is done in different steps and part of the estimation is already done in a first model, delivering a certain level where a second model builds further upon. The first model is then included in the second model using offset variables.
For example for the Poisson case, we have

\[ \lambda_i = \exp(\text{score}_i) = \exp(\beta_0) \exp(\sum_{j=1}^{p} \beta_j x_{ij}) \exp(\text{offset}_i) \]

So when the offset is a time period, one uses \( \text{offset}_i = \ln(t_i) \) so that

\[ \frac{\lambda_i}{t_i} = \exp(\beta_0) \exp(\sum_{j=1}^{p} \beta_j x_{ij}). \]

This becomes more clear when we illustrate this with a Poisson model that contains only one explanatory variable \( x \), which has \( m \) levels - for each level \( x_{ij} (1 \leq j \leq m) \) is the indicator that equals 1 if observation \( i \) belongs to category \( j \) or else 0. If we leave the intercept out, so \( \beta_0 = 0 \), then we can immediately prove that the ML estimators equal the mean of their category, \textit{weighted} with this offset variable \( t_i \) per category. First, note that for every observation of level \( j \) :

\[ \lambda_i = \mathbb{E}[Y_i] = \exp(\text{score}_i) = \exp(\sum_{j=1}^{m} \beta_j x_{ij}) \exp(\text{offset}_i) = t_i \exp \beta_j. \]

Second, the ML equations (2.11) have to be fulfilled (put \( w_i = 1 \)):

\[
\sum_{i=1}^{n} x_{ij} y_i = \sum_{i=1}^{n} x_{ij} \lambda_i \\
\Leftrightarrow \sum_{i|x_{ij}=1} y_i = \sum_{i|x_{ij}=1} \lambda_i = \sum_{i|x_{ij}=1} t_i \exp \beta_j \\
\Leftrightarrow \frac{\sum_{i|x_{ij}=1} y_i}{\sum_{i|x_{ij}=1} t_i} = \exp \beta_j. \tag{2.15}
\]

2.9 Weight

As we proved in (1.3.2), the variance is divided by the exposure; in GLM this means that the exposure should always be inserted in the model as weights. In that case, the dispersion parameter \( \phi \) is replaced with \( \phi/w_i \) for observation \( i \) with weight \( w_i \). Hence one should carefully interpret the estimated outcomes. We use the exposure of each observation to estimate the coefficients \( \beta_i \) of the model, so the exposure does change the score function, but when we determine the expected claim frequency as in (2.10), this has to be interpreted on an annual basis. So when the exposure period \( w_i \) is not one year, the expected outcome is \( w_i \lambda_i \).
This second type of correction, in comparison to the first one, the offset, delivers different estimates since there is a subtle yet very important difference. This becomes more clear when we retake the Poisson model that contains only one explanatory variable $x$, which has $m$ levels. If we leave the intercept out, so $\beta_0 = 0$, then we can immediately prove that the ML estimators equal the weighted mean of the observations of this category, but with the weight $t_i$ taken into account per observation (compare with the offset in the next section). First, note that for every observation of level $j$:

$$\lambda_i = E[Y_i] = \exp(\text{score}_i) = \exp(\sum_{j=1}^{m} \beta_j x_{ij}) = \exp \beta_j.$$ 

Second, the ML equations (2.11) have to be fulfilled:

$$\sum_{i=1}^{n} w_i y_i x_{ij} = \sum_{i=1}^{n} w_i \lambda_i x_{ij}$$

$$\Leftrightarrow \sum_{i|x_{ij}=1} w_i y_i = \sum_{i|x_{ij}=1} w_i \lambda_i = \sum_{i|x_{ij}=1} w_i \exp \beta_j$$

$$\Leftrightarrow \frac{\sum_{i|x_{ij}=1} w_i y_i}{\sum_{i|x_{ij}=1} w_i} = \exp \beta_j.$$ (2.16)

So when we want to estimate the claim frequency, and this has to confirm of course the empirical frequencies as in (2.15), one has to use as $y$ variable the ratio $\text{number of claims}/\text{exposure period}$. Then we immediately have

$$\sum_{i|x_{ij}=1} w_i \frac{n_i}{w_i} = \sum_{i|x_{ij}=1} w_i \lambda_i = \sum_{i|x_{ij}=1} w_i \exp \beta_j$$

$$\Leftrightarrow \frac{\sum_{i|x_{ij}=1} n_i}{\sum_{i|x_{ij}=1} w_i} = \exp \beta_j.$$ (2.17)

which is exactly the desired observed claim frequency. This is the correct claim frequency in opposite to the one above (2.16). This could be misleading since modelling $y = n_{\text{claim}}$ with weight $\text{dur}$ would yield lower AIC values - hence a better model? No, this is because the response values would seem to be closer to each other. This can be for example seen with the likelihood function. In $L$ as in (2.3), we would use for $y_i$ the netto number of claims - watch the first term. These would be smaller values than when $y_i$ would be the ratio $\text{number of claims}/\text{exposure period}$, since these are the exact same numbers, number of claims, but now first divided by a number smaller than 1. So larger values would be produced, leading to more variation in the response so automatically lower AIC score since the model is harder to fit. Finally, we want to stress the fact that we here proceed with the
weighted model instead of the offset model. Both estimate perfectly the correct empirical frequencies, but the weights have to appear in the variance function, which is not the case for the offset model. Moreover the AIC values are lower for the weighted model.

Note that the same reasoning applies for the gamma distribution, if we also use the ln link function, since the ML equations (2.14) are the same except for the denominator $\mu_i$, which can be multiplied away since it appears in the whole equation. So the claim severity model with weight $n_{claim}$ and one categorical variable gives also as estimated means

$$\sum_{i \mid x_{ij} = 1} w_i \frac{costult_i}{n_{claim_i}} = \sum_{i \mid x_{ij} = 1} w_i \lambda_i = \sum_{i \mid x_{ij} = 1} w_i \exp \beta_j$$

$$\sum_{i \mid x_{ij} = 1} \frac{costult_i}{n_{claim_i}} = \exp \beta_j.$$

### 2.10 Tariff

As explained in the first section, the pure premium is the expected cost the insurer will lose from policyholders’ claims and the technical tariff, developed for each policyholder, tries to estimate this. As the policyholders are divided into categories, the tariff consists of tariff cells: policyholders from the same category are in the same tariff cell and will have to pay the same premium. This is only true in theory, since the market premium however will differ from this technical premium for several reasons. First of all the size or growth of the insurer is important since it indicates its influence on and position in the market. Secondly, the tariff of the competitors cannot be much lower or one would lose too many clients. Thirdly it is clear that if bad accidents with a lot of damage happened in the past, the associated pure premium would be devastatingly high to pay for an insurance contract. And vice versa, policyholders who never produced a claim would get a free insurance! This will be discussed in detail in chapters six and seven.

### 2.11 Tweedie models

Within the exponential dispersion family, a (family of) probability distribution(s) is uniquely characterized by its variance function [5]. So far we have encountered, disregarding weights, that the variance is equal to the mean (Poisson distribution) or proportional to the squared mean (gamma distribution). Moreover the variance plays an important role with respect to scale invariant distributions. Scale invariance of a random variable $Y$ is achieved if, for
every $c > 0$, the distribution of $cY$ (dilatation of $Y$) belongs to the same family as the distribution of $Y$. This is of course desired here, for all ratios in (1.1) since the tariff should be independent of the used currency, where the transformation to another currency is the result of multiplying by a positive number; also the claim frequency may be measured in per cent or per mille [5]. It turns out that the Tweedie models, whose variance function or variance satisfies

$$V(\mu) = a\mu^p \text{ or } \text{Var}(Y) = \frac{\phi}{w}\mu^p$$

with $\mu = E[Y]$ and $a > 0$,

are the only scale invariant exponential dispersion models (this relation is known in physics as fluctuation scaling).

We find thus all kinds of distributions (discrete, continuous, mixed) in the Tweedie class: many applications (not in tariff models) for $p < 0$ (continuous distributions on whole real axis) and $p = 0$ (the normal distribution),

no distributions exist for $0 < p < 1$,

$p = 1$ (Poisson distribution) is often used to model claim frequency,

often used to model pure premium is $1 < p < 2$ (the compound Poisson distribution with gamma distributed claims),

and $p \geq 2$ often used to model claim severity ($p = 2$ the gamma distribution, $p = 3$ the inverse Gaussian distribution, and $2 < p < 3$, $p > 3$ for continuous positive distributions).

The name of this family refers to the first man (apparently) that studied the compound Poisson model with gamma severities in the context of the exponential family origintweedie.

In [6] the cumulant function is calculated for $1 < p < 2$ and $p > 2$:

$$b(\theta) = -\frac{1}{p-2} \left[ -(p-1)\theta \right]^{(p-2)/(p-1)}$$

and we already know that $b(\theta) = \exp(\theta)$ for the Poisson ($p = 1$) and $b(\theta) = -\ln(-\theta)$ for the gamma distribution ($p = 2$). Moreover the relations hold:

$$b(\mu) = \frac{\mu^{2-p}}{2-p} \text{ for } p \neq 2, \text{ otherwise } b(\mu) = \ln \mu,$$

$$\theta(\mu) = \frac{\mu^{1-p}}{1-p} \text{ for } p \neq 1, \text{ otherwise } \theta(\mu) = \ln \mu.$$

### Compound Poisson distribution

Except for the special cases $p \in (0, 1, 2, 3)$, the density of the Tweedie distribution involves an infinite sum. We will only consider here now the case $1 < p < 2$: the compound
Poisson, also called Poisson-gamma or compound gamma distributions since this is the sum of gamma distributed values (the losses), for a number that is Poisson distributed (the number of claims). These are thus mixed distributions with mass at zero and support on the positive real values.

The full loglikelihood is given by

\[
\mathcal{L}(y_i \neq 0) = \sum_{i=1}^{n} \ln(V_i) + \frac{w_i}{\phi} \left[ y_i \mu_i^{1-p} - \frac{\mu_i^{2-p}}{2-p} \right] - \ln(y_i),
\]

\[
\mathcal{L}(y_i = 0) = -\sum_{i=1}^{n} \frac{w_i \mu_i^{2-p}}{\phi (2-p)}
\]

with \( V_i = \sum_{j=1}^{\infty} V_{ij} = \sum_{j=1}^{\infty} \frac{w_i^{1-j(1-p)} y_i^{-j^{2-p}}}{\phi^{j(1-p)} (2-p) j! \Gamma(-j(1-p))} \).

This loglikelihood can indeed be compared with the general density formula for a member of the exponential family:

\[
f_{Y_i}(y_i; \theta_i, \phi) = \exp \left( \frac{y_i \theta_i - b(\theta_i)}{\phi / w_i} + c(y_i, \phi, w_i) \right)
\]

\[
= \exp \left( \frac{w_i}{\phi} \left( y_i \theta_i + \frac{1}{p-2} \left[ -(p-1) \theta (p-2)/(p-1) \right] \right) + c(y_i, \phi, w_i) \right)
\]

\[
= \exp \left( \frac{w_i}{\phi} \left( y_i \mu_i^{1-p} - \frac{\mu_i^{2-p}}{2-p} \right) \right) \exp(c(y_i, \phi, w_i))
\]

\[
= \exp \left( \frac{w_i}{\phi} \left( y_i \mu_i^{1-p} - \frac{\mu_i^{2-p}}{2-p} \right) \right) V_i / y_i
\]

so that with this parameterization,

\[
\exp(c(y_i, \phi, w_i)) = \frac{V_i}{y_i}.
\]

We can take a closer look at the link between the different distributions involved, as done in [11]. Suppose that the number of claims is Poisson distributed with mean \( \lambda \), and the severity is gamma distributed with index (or shape) parameter \( \alpha \) and scale parameter \( \beta \). The compound distribution can be parameterized with \( \phi \), \( \mu \) and \( p \) where then

\[
\mu = \lambda \frac{\alpha}{\beta} \quad , \quad p = \frac{\alpha + 2}{\alpha + 1} \quad , \quad \phi = \frac{\lambda^{1-p}}{(\alpha/\beta)^{2-p}}.
\]
The probability of being zero is then
\[ p_0 = \exp(-\lambda) = \exp\left(-\frac{\mu^{2-p}}{\phi(2-p)}\right). \]

**ML equations, Fischer information matrix**

The expression for the variance function, in combination with a multiplicative model - the 
\( g = \ln \) link function, allows us to calculate the ML equations (note that (2.14) is confirmed again):
\[
\sum_{i=1}^{n} w_i \frac{y_i - \mu_i}{\mu_i^{p-1}} x_{ij} = 0 \quad (1 \leq j \leq p)
\]
with the connection
\[ \ln \mu_i = \sum_{j=1}^{p} x_{ij} \beta_j. \]

Note that when using an offset \( z_i = \ln(u_i) \), so
\[ \ln \mu_i = z_i + \sum_{j=1}^{p} x_{ij} \beta_j \iff \ln \frac{\mu_i}{u_i} = \sum_{j=1}^{p} x_{ij} \beta_j, \]
the ML equations become \[5\]
\[
\sum_{i=1}^{n} w_i \frac{y_i - \mu_i u_i}{(\mu_i u_i)^{p-1}} x_{ij} = 0 \iff \sum_{i=1}^{n} w_i u_i^{2-p} \frac{y_i - \mu_i}{\mu_i^{p-1}} x_{ij} = 0 \quad (1 \leq j \leq p).
\]

So using the offset \( \ln(u_i) \) is equivalent to using a GLM with observations \( y_i/u_i \) and weights \( w_i u_i^{2-p} \).

The Fisher information matrix (2.5) becomes here for the Tweedie models with log link:
\[ I_{jk} = \sum_i \frac{w_i}{\phi \mu_i^{p-2}} x_{ij} x_{ik}. \]

**In practice**

As already mentioned, tweedie models with \( 1 < p < 2 \) can be used to analyse the pure premium directly, however this is not a common strategy \[5\]. Because claim frequency is often more stable than claim severity hence provides more correct estimators, and because of the transparency and insight in rating factors, the standard method involves a separate analysis for both. Multiplying the results then gives an estimate for the pure premium; this will be performed on the dataset.
2.12 Goodness of fit measures

The following three criteria are well known and will be useful in selecting among models, with smaller values representing better model fit.

The Akaike information criterion (AIC) is a measure that balances model fit against model simplicity: it takes into account the loglikelihood $L$ and the number of parameters $r$ (don’t forget the scale parameter if it is also estimated). AIC has the form

$$AIC = -2L + 2r,$$

so it penalizes overfitting - using too much parameters. An alternative form is the corrected AIC given by

$$AICC = -2L + 2r \frac{n}{n - r - 1},$$

where $n$ is the total number of observations used and clearly converges to AIC for large $n$ and small $r$. A third, similar measure is the Bayesian information criterion (BIC), which is bigger than the AIC (or AICC) for large enough $n$:

$$BIC = -2L + r \ln(n).$$

A fourth criteria, the consistent AIC, is not given in SAS and less used:

$$CAIC = -2L + r(\ln(n) + 1).$$

Deviance

Another statistic that is often used to compare models, but has also meaning on its own for a specific model, is the deviance. This is defined as the difference in loglikelihood of the saturated model and the model under consideration:

$$D = 2\phi \left[ L(y_i; y_i) - L(\hat{\mu}_i; y_i) \right].$$

Note that the deviance also has a limiting chi-square distribution; as the Pearson statistic it can be used to estimate $\phi$ but as we already noted we will use the Pearson estimator because it appears to be more robust.

We can calculate this for the Poisson distribution (with (2.9) and $\phi = 1$):

$$D = 2\phi \sum_{i=1}^{n} \left[ w_i \left( y_i \ln(\lambda_i) - \lambda_i \right) \pm w_i y_i \ln(w_i) \mp \ln((w_i y_i)!) - w_i \left( y_i \ln(y_i) - y_i \right) \right]$$

$$= 2 \sum_{i=1}^{n} w_i \left[ y_i \ln \left( \frac{\lambda_i}{y_i} \right) - \lambda_i + y_i \right].$$
And for the gamma distribution we find (with (2.13)):

\[ D = 2\phi \sum_{i=1}^{n} \frac{w_i}{\phi} \left( -\frac{y_i}{y_i} - \ln(y_i) \right) \mp \ln\left( \Gamma\left( \frac{w_i}{\phi} \right) \right) \pm \ln\left( \frac{w_i y_i}{\phi} \right) \mp \ln(y_i) - \frac{w_i}{\phi} \left( -\frac{y_i}{\mu_i} - \ln(\mu_i) \right) \]

\[ = 2 \sum_{i=1}^{n} w_i \left[ \frac{y_i - \mu_i}{\mu_i} - \ln\left( \frac{y_i}{\mu_i} \right) \right]. \]

Finally for the Tweedie distribution we see that the infinite sum \( V_{i} \) in (2.18) disappears:

\[ D = 2\phi \sum_{i=1}^{n} \left[ \pm \ln(V_i) + \frac{w_i}{\phi} \left( \frac{y_i^{2-p}}{1-p} - \frac{y_i^{2-p}}{2-p} \right) - \frac{w_i}{\phi} \left( y_i \frac{\mu_i^{1-p}}{1-p} - \frac{\mu_i^{2-p}}{2-p} \right) \mp \ln(y_i) \right] \]

\[ = 2 \sum_{i=1}^{n} w_i \left[ \frac{y_i^{2-p}}{(1-p)(2-p)} - y_i \frac{\mu_i^{1-p}}{1-p} + \frac{\mu_i^{2-p}}{2-p} \right]. \]

We will see in chapter six that modelling the Tweedie distribution is possible with this deviance only.
Chapter 3

Credibility theory

When classifying the population (policyholders) into different risk categories, the data is often not equally divided leading to risk cells with far more or less observations than other cells. Especially when a risk factor is the product of categorical variables, this occurs even when you have a lot of data at your disposal. A possibility for handling a lack of data, may be to merge different risk cells together; however when the categorical variable is not ordinal but nominal this may lead to strange constructions (for example merging age class [18, 20] with [21, 23] makes sense whereas merging geographical zones 1 and 2 doesn’t mean much). We adopt the notation from [5] to denote such a nominal categorical variable a multi-level factor (MLF). A classical example is the car model: thousands of different types and brands of cars exist, among which there is no obvious ordering. The same for a geographical division with postcodes. Even the policyholder itself can be viewed as a MLF which results in his BM (experience rating). Problems of this kind, handling a lack of data on some levels/variables, are studied in credibility theory.

3.1 Credibility estimator

Suppose we are interested in a ratio $Y = X/w$ and a MLF with $k$ levels leading to subpopulations $Y_1, \ldots, Y_k$. The observations are denoted by $Y_{ij}$ with weights $w_{ij}$ where the first subscript $i$ refers to the level of the MLF and the second subscript $j$ refers to the observation number (within the subpopulation). The average of ratio $Y$ of the subpopulation is then

$$\bar{Y}_j = \frac{1}{\sum_i w_{jt}} \sum_i w_{jt} Y_{jt} = \frac{1}{\sum_i w_{jt}} \sum_i w_{jt} X_{jt}$$
where the whole population average $\mu$ of this ratio may for example be estimated by

$$\hat{\mu} = \bar{Y} = \frac{1}{w} \sum_{j=1}^{k} w_j \bar{Y}_j = \frac{1}{w} \sum_{j=1}^{k} \sum_{l} w_{jl} Y_{jl} = \frac{1}{w} \sum_{j=1}^{k} \sum_{l} X_{jl} w_j, \quad w = \sum_{j=1}^{k} w_j.$$

The idea of credibility theory is then to create more stable estimators for the cells with a lack of data, with help of $\mu$: the **credibility estimator** is a combination of both

$$\kappa_j \bar{Y}_j + (1 - \kappa_j)\mu$$

where $0 \leq \kappa_j \leq 1$ is called the **credibility factor** or weight. Some important remarks follow immediately:

- the more data subpopulation $j$ contains, the larger $\kappa_j$ should be,
- the less variation there is in subpopulation $j$, the larger $\kappa_j$ should be,
- the more variation between the subpopulations $Y_i$ and $Y_j$, the larger $\kappa_j$ should be ($\mu$ says little about the subpopulations).

Note that nowadays, random parameters are often used to derive credibility estimators. For example the expected number of claims for a certain policyholder $\lambda_i$ is the outcome of a random variable $\Lambda_i$, which are assumed to be independent and identically distributed among the subpopulations (see further).

### 3.2 Bühlmann-Straub model

I adopt here the notations in [5].

For the Bühlmann-Straub model, only one MLF is considered in the model without any other factors. Moreover, only the mean and variance are assumed to be known - which leaves more freedom thus is an advantage in applications. From now on we assume that the first two moments of all considered random variables exist. Consider the relativity $u_j$ of subpopulation $j$, this is assumed to be an observation of a random effect $U_j$:

$$E [Y_{ij} | U_i] = \mu U_i \text{ with } E [U_i] = 1.$$  \hfill (3.1)

So in GLM the relativities $u_j$ are the factor in which that this subpopulation’s ratio differs from the whole population’s ratio. Expressed in the variable $V_j := \mu U_j$ we have then

$$E [Y_{ij} | V] = V_i \text{ with } E [V_i] = \mu \text{ and } Var [Y_{ij} | V_i] = \frac{\phi}{w_{ij}} V_i^p.$$
This assumption about the variance is motivated by the fact that in GLM, we assumed that the data followed a Tweedie model (with dispersion $\phi$). Assuming that the $V_i$ are identically distributed with mean $\mu > 0$ and variance $\tau^2 > 0$, we denote $\sigma^2 = \phi E[V_i^p]$ so that

$$E[\text{Var}(Y_{ij}|V_i)] = \frac{\sigma^2}{w_{ij}}.$$  
(3.2)

Note that these assumptions imply that

$$\text{Var}(Y_{ij}) = \text{Var}[E(Y_{ij}|V_i)] + E[\text{Var}(Y_{ij}|V_i)] = \tau^2 + \frac{\sigma^2}{w_{ij}}.$$ 

The credibility estimator of a random effect $V_i$ is defined as the linear function $\hat{V}_i$ of the observations $Y$ that minimize the mean square error

$$E[(h(Y) - V_i)^2]$$

among all linear functions $h(Y)$, $\hat{V}_i$ is sometimes also called the best linear predictor. The solution is the estimator of Büllmann and Straub:

$$\hat{V}_i = \kappa_i \bar{Y}_i + (1 - \kappa_i)\mu$$

with $\kappa_i = \frac{w_i}{w_i + \sigma^2/\tau^2}$ (0 $\leq \kappa_i \leq k$).

Proof can be found in [3].

Since the observations $Y_{ij}$ and their weights $w_{ij}$ are given, it is clear that we only need $\mu, \sigma^2$ and $\tau^2$ to estimate the random effect $V_i$. Because of (3.2), $\sigma^2$ can be seen as the within-group variance, so the variance in a certain subpopulation, so that

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{k} (n_i - 1)\hat{\sigma}_i^2}{\sum_{i=1}^{k} (n_i - 1)}$$

with $\hat{\sigma}_i^2 = \frac{1}{n_i - 1} \sum_j w_{ij}(Y_{ij} - \bar{Y}_i)^2$

where $n_i$ is the number of observations in subpopulation $i$. From $\tau^2 = \text{Var}(V_i)$ it follows that $\tau^2$ is rather the between-group variance, so the variances between the $k$ subpopulations:

$$\hat{\tau}^2 = \frac{\sum_{i=1}^{k} w_i(\bar{Y}_i - \bar{Y})^2 - (k - 1)\hat{\sigma}^2}{w - \sum_{i=1}^{k} w_i^2/w}.$$ 

It turns out that these are both unbiased estimators (under the made assumptions).
3.3 Combine GLM with credibility theory

Often other risk factors or explanatory variables are used beside the MLF, denote the relativities of these \( r \) other variables with \( \gamma_1, \ldots, \gamma_r \) for risk cell \( i \). Then (3.1) is generalized to

\[
E \left[ Y_{ijk} | U_j \right] = \mu \gamma_i^1 \ldots \gamma_i^r U_j \quad \text{with} \quad E[U_i] = 1.
\]

Thus we have three subscripts for an observation: the first refers to the cell based on the \( r \) variables (non-MLF), the second to the subpopulation with respect to the MLF (or random effect \( U_j \)), the third to the observation itself. Assume we know \( \mu \) and \( \gamma_i := \gamma_i^1 \ldots \gamma_i^r \), then we want to compute an estimator of \( U_j \) or \( V_j = \mu U_j \). We still assume that the \( p \) subpopulations, hence random vectors \( (Y_{ijk}, V_j) \), are independent and that the \( p \) random effects \( V_j \) are identically distributed with mean \( \mu > 0 \) and variance \( \tau^2 > 0 \). The \( Y_{ijk} | V_j \) are indendent and follow a Tweedie distribution in GLM thus it is natural to assume also that

\[
E \left[ Y_{ijk} | V_j \right] = \gamma_i V_j \quad \text{with} \quad Var \left[ Y_{ijk} | V_j \right] = \frac{\phi}{w_{ijk}} (\gamma_i V_j)^p.
\]

We also generalize (3.2):

\[
E \left[ \text{Var} \left( Y_{ijk} | V_j \right) \right] = \frac{\gamma_i^p \sigma^2}{w_{ijk}}.
\]

So in order to reduce this problem to the classical Büllmann-Straub problem treated in the first section, we transform the observations:

\[
\tilde{Y}_{ijk} = \frac{Y_{ijk}}{\gamma_i}, \quad \tilde{w}_{ijk} = w_{ijk} \gamma_i^{-2+p}
\]

so that

\[
E \left[ \tilde{Y}_{ijk} | V_j \right] = V_j, \quad Var \left[ \tilde{Y}_{ijk} | V_j \right] = \frac{\phi}{w_{ijk}} V_j^p, \quad E \left[ \text{Var} \left( \tilde{Y}_{ijk} | V_j \right) \right] = \frac{\sigma^2}{\tilde{w}_{ijk}}.
\]

So we apply the results to get immediately that

\[
\hat{V}_j = \kappa_j \bar{Y}_j + (1 - \kappa_j) \mu \quad \text{with} \quad \kappa_j = \frac{\tilde{w}_i}{\tilde{w}_i + \sigma^2 / \tau^2}
\]

and

\[
\bar{\tilde{Y}}_j = \frac{\sum_{i,k} \tilde{w}_{ijk} \tilde{Y}_{ijk}}{\tilde{w}_j} = \frac{\sum_{i,k} \tilde{w}_{ijk} Y_{ijk} \gamma_i^{1-p}}{\sum_{i,k} \tilde{w}_{ijk} \gamma_i^{2-p}}.
\]

The estimators for \( \sigma^2 \) and \( \tau^2 \) are easily generalized, note that these are no longer unbiased if we estimate the \( \gamma_i \) by GLM:

\[
\hat{\sigma}^2 = \frac{1}{n_j - 1} \sum_{i,k} \tilde{w}_{ijk} (\bar{\tilde{Y}}_{ijk} - \bar{\tilde{Y}}_j)^2
\]

\[
\hat{\tau}^2 = \frac{\sum_{j=1}^p \tilde{w}_j (\bar{\tilde{Y}}_j - \bar{\tilde{Y}})^2 - (p - 1) \hat{\sigma}^2}{\tilde{w} - \sum_{j=1}^p \tilde{w}_j^2 / \tilde{w}}.
\]
Chapter 4

Dataset

This dataset contains data from a certain insurance company which will remain anonymous. The data represent car insurance policies (claims and costs with respect to third party liability only), spread over several years (the insured periods take place in ). It deals with policies where the insured is a person, a few small businesses or companies are also involved. Moreover some data cleaning is already done: policies without known claim history and policies with a negative claim severity are excluded.

4.1 Variables: overview

4.1.1 A priori variables

Identification variables

The policynumber is at our disposal but will not be used in our analysis.

Scale variables

Probabilities and frequencies are always computed in function of dur, which is the time length of the insured period, where dur = 1 corresponds with one full year. This variable is used as a scale weight variable since the claim data have to be correctly interpreted according to the duration of the insured period during which the accident happened (see section [1.3]).

Premium is the paid premium for the insured period dur, so clearly already corrected or weighted for the according duration. This variable is also used as a scale weight variable when modelling the loss ratio.
Continuous variables

Age is the age of the policyholder.
Pow is the power of the car (in kW).
Budget is the discount that was granted when the insurance was closed.

Categorical variables

Age_CLASS_CL is the age class as developed by the company. This is not correctly included in the datasystem: many overlappings occur and misplaced ages although then their frequency is almost zero. We will try to improve this segmentation.

Geocode is a code of geographical zones, it is however not very useful since the division is too precise with over 1100 levels.

Zone_CL is a geographical categorization (5 zones), developed in the 90s based on descriptive statistics and claim frequencies. The lower, the more safe the zone should be (less claims). This is however informative and is replaced in the tarification by the following segmentation.

Zone is a new geographical categorization (10 zones), developed a few years back. It is based on statistics of the postal areas combined with ZoneG. Because of the small number of observations in zone 0 (the safest zone), the zones 0 and 1 will be put together leading to the variable newzone.

ZoneG is a categorization that combines geographical data with the social-economic and professional environment of the policyholder; it has 35 levels.

Cover describes the type of coverage that was insured: Full omnium, Limited omnium, TPL and Other. Only claims with respect to civil responsibility are described here (number and costs) but this may off course have happened in a policy where the original coverage was more extended than this.

Pow_cat is the categorical division according to the power in 8 categories (\([0, 40[\), \([40, 50[\), \([50, 60[\), \(\ldots\), \([100+]\).

Brand is the brand of the car.

Categorical variables that may be used as continuous variables

Age_policy is the age of the policy, limited to five years.
Age_car is the age of the car. Since numbers up to 100 have been registered here, this age was limited to thirteen years.
4.1 Variables: overview

**Capacity** is the number of persons that can be transported by the vehicle (range from 2 (or smaller) to 9).

**Npol** indicates how much different insurance types the policyholder has currently closed with this insurer. This varies from 1 to 4 since only the four main categories car (TPL), life, fire and family insurance are considered.

**Indicator variables**

**Code_sport** indicates whether the car is characterized as a sportscar yes (1) or no (0).

**Overtake** indicates whether the policyholder was insured elsewhere before this contract or not.

**Ind_claim_5year** indicates whether there has been a claim in the five years preceding the start date of the contract, where the policyholder was declared in fault.

**Diesel** indicates whether the car uses diesel (1) or petrol (0) as fuel type.

**Private** equals 1 if the car is used for private purposes, in contrary to jobrelated transport (0).

**Ind_reduction** indicates whether there was a discount applied or not.

**Ind_omnium** indicates whether the coverage includes an omnium formula (1 for Full omnium or Limited omnium) or not (0).

**Sex** is the sex of the policyholder (0 or unknown for small bussinesses), this variable may not directly be used however for tarification (by law against discrimination).

4.1.2 A posteriori variables

**Nclaim** is the number of claims where the policyholder was in fault, during the insured period. This variable is also used as a weight variable when modelling the claim severity.

**Cost** is the sum of payments and reserve minus the reclaims.

**Costlim** is the cost limited to euro.

**Costult** is the estimated expected total cost, which is the limited cost multiplicated by the Best Estimate coefficient of that year (calculated by IBNR triangles). For all large claims, the part above euros, is taken in consideration for the calculation of this BE. In this way these upper parts are spread over the population and divergent data are avoided. This costult consists thus of two parts: one part related to this particular observation and a second part representing a mean cost coming from these large claims over several years.

**BM** is the bonusmalus of the policyholder, we already explained why this doesn’t correspond with the true BM of the policyholder, see section 1.2. A correction based on the age of the policyholder and the year this person has obtained his driving license, is possible.
4.2 Univariate statistics and first insights

In a first step, a univariate analysis is performed to explain the different levels of the variables and their meaning. Since the goal will be to predict the posteriori variables such as number of claims and the corresponding costs in order to determine a premium, the focus will be on the distribution of these among the various levels and more specifically even for nclaim. This along with the duration since this measures the representability and credibility of these levels; the duration is of importance here rather then the number of observations!

In our frequency model, i.e. Poisson model for the number of claims, we will use dur as weight and we will model in fact the variable nclaim/dur, so the claim frequency, as explained in (2.17). Without intercept and with only one explanatory variable we can determine the weighted means and confidence intervals; recall that

\[
\exp \beta_j = \frac{\sum_{i|x_{ij}=1} nclaim_i}{\sum_{i|x_{ij}=1} dur_i}.
\]

Since we are interested in the relativities, the exponential of these estimators, we will compute the 95% confidence interval with the exponential of these boundaries:

\[
L = \exp(\hat{\beta}_j - 1.96\sqrt{b_{jj}}), U = \exp(\hat{\beta}_j + 1.96\sqrt{b_{jj}}).
\]

The same can be done to model the claim severity (severity model); this will be done with a gamma model where now nclaim is the weight and the ratio costult/nclaim is modelled. Note that the weight is thus zero for all policies where no claim was reported; hence the model uses the reduced dataset which contains only the claims. This is important to distinguish the pure premium and claim severity! In the following tables, we always give the PctSum, which denotes the share (in percentage) of a certain level of a variable, in the total sum of the duration. Then several means are given, first of all the yearly means, denoted by \(\text{mean } y\): the mean of the considered variable \(x\) with respect to the duration, so \(\sum x_i / \sum dur_i\) for a certain level.

- Mean \(y\) premium is then the average yearly premium since this gives \(\sum \text{premium}_i / \sum dur_i\).
• Mean $y \text{costult}$ is then exactly the pure premium, the ultimate cost per year, since this gives $\sum \text{costult}_i / \sum \text{dur}_i$.

• Mean $y \text{nclaim}$ is then exactly the claim frequency since this gives $\sum \text{nclaim}_i / \sum \text{dur}_i$.

Finally the average claim severity is given by mean $n \text{costult}$: this is $\sum \text{costult}_i / \sum \text{nclaim}_i$. But this is only calculated for the policies where minimum one claim was reported! In opposite to the pure premium, which involves also the duration of the contracts where no claim was registered. The correct interpretation is that the claim severity is the average cost of a claim, knowing that a claim occurred; whereas the pure premium is the average cost of a policy, regardless we know whether a claim was reported or not.

These values can be calculated by adding all statistics together for different levels of a variable, or they can be obtained as estimates in a Poisson model (for the claim frequency) or a gamma model (for the claim severity). In (2.17) we showed that this indeed gives identical results; also the standard error and CI are given which are of course valuable information. The figures for claim frequency and claim severity in this chapter are the result of these models; CI limits (upper $U$ and lower $L$ as defined in (4.1)) are always represented by a dotted line. For the pure premium, no model was used because first of all, this would involve a Tweedie model where the parameter $1 < p < 2$ should be estimated from the data, and second, in practice it is in most cases better (more robust estimates) to obtain pure premium estimates from multiplying the estimates from frequencies and severities. The CI is therefore constructed with the sum of the variances as explained in (2.7) ($F$ stands for frequency, $S$ for severity, and $\beta$ denotes here the absolute estimates, so before taking exp):

$$\left[ \exp\left( \hat{\beta}_j^F + \hat{\beta}_j^S - 1.96 \sqrt{\text{Var}\left[ \hat{\beta}_j^F \right]} + \text{Var}\left[ \hat{\beta}_j^S | \text{nclaim} \right] \right), \exp\left( \ldots + 1.96 \sqrt{\ldots} \right) \right].$$

This third figure thus always contains the pure premium estimate and CI resulting from the Poisson and gamma model, and the average premium as simply calculated from the data (to show possible mispricing or confirmation of the current premium height).

The estimated means in the frequency model are always exactly correct with the empirical frequencies, calculated from the dataset. In the severity, thus also the pure premium model, this is not the case. The reason for this are the claims with zero ultimate cost; we choose not to exclude these from our frequency analysis because a claim was reported so to correctly estimate the frequency, these should be included. In the next table one can
see that it covers \( n \) claims, \( n \) pct of the total number of claims. Only small residuals 
\((= \text{estimated claim severity} - \text{empirical claim severity (including zero costs)})\) are observed 
(typically a deviation of about \( \pm 1\% \) relative to the estimate).

Note that we will not cover every explanatory variable in a thorough way - since there 
are 24 we would drown in unnecessary figures and tables. Only those which will turn out 
to be important will be covered - these tables will be mainly used to gain a first insight 
and will be referred to when drawing conclusions from models later on.

### 4.2.1 Overall view

**Table 4.1:** Key variables statistics over the whole dataset, and the part with non-zero costult.

<table>
<thead>
<tr>
<th></th>
<th>dur</th>
<th>nclaim</th>
<th>premium</th>
<th>costult</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sum</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean ( y )</td>
<td></td>
<td>(claim frequency)</td>
<td></td>
<td>(pure premium)</td>
</tr>
<tr>
<td>Mean ( n )</td>
<td></td>
<td></td>
<td></td>
<td>(claim severity)</td>
</tr>
<tr>
<td>Sum</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean ( y )</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean ( n )</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The observed number of claims for one policy goes from 0 to 3, \( n \) pct of the duration 
represents policies where no claims were observed! Of the \( n \) observations, only \( n \) policies reported 2 claims and another \( n \) reported 3 claims. Notice that the total duration 
sum is equivalent to about \( n \) years. The mean is calculated with respect to the 
duration: for the number of claims for example, this is the average \( n \) for a duration of 
one year since \( n \). Note that the lossratio is only \( n \) \%, this is lower 
because of the data cleaning already done before the analysis; the policies without known 
claim history, so the most risky clients actually, were omitted. So new clients are of course 
included, but those clients that have driven a car in the past, thus were already insured in 
the past, but did not give any records or proof of this insurance, are not included.

To illustrate the difference for a distribution analysis with and without weights, we 
performed both for the premium resulting in two histograms (figure 4.1). Clearly, the 
weighted approach has smaller counts, since an observation does not necessarily counts 
for 1 (only if the weight \( \text{dur} \) equals 1). Also the weighted histogram allows a continuous 
approach from zero, where the non-weighted histogram is discontinuous for small values.
4.2 Univariate statistics and first insights

Figure 4.1: The difference of a non-weighted (left figure) versus a weighted (right figure) analysis for premium.

Dataset claims

To give an idea of the ultimate costs arising from a claim, a frequency analysis is also performed for costult, only for the part of the dataset that generated claims (so all zero-costs are excluded) (figure 4.2). Note that this analysis is weighted with nclaim.

Figure 4.2: Frequency histogram of the ultimate cost.
4.2.2 Age of the vehicle, policy and policyholder

Age of the vehicle

For age of the vehicle, values from 0 to 100 are observed. Since only \( \text{pct} \) of the duration is represented by values of 18 or higher, and mostly because of the unreliability of higher values, this variable is cut off at 18: level 18 contains all cars of this age or older. So the range 0 – 18 is displayed on figure 4.3. Recall that \( L \) is the lower and \( U \) the upper limit of the confidence interval.

\[\text{fig1agecarlim.JPG}\]

**Figure 4.3:** Average claim frequency and 95% CI (dotted lines) for the age of the vehicle.

\[\text{fig2agecarlim.JPG}\]

**Figure 4.4:** Average claim severity and 95% CI (dotted lines) for the age of the vehicle.
Figure 4.5: Average pure premium and 95% CI (dotted lines) for the age of the vehicle.

Table 4.2: Frequency of the levels of age_car with respect to the duration.

<table>
<thead>
<tr>
<th>Age_car</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>PctSum dur</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age_car</td>
<td>10</td>
<td>11</td>
<td>12</td>
<td>13</td>
<td>14</td>
<td>15</td>
<td>16</td>
<td>17</td>
<td>18(+)</td>
<td></td>
</tr>
<tr>
<td>PctSum dur</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

We see that every level up to 9 represents about to pct of the dataset. The claim frequency, severity and pure premium with the age of the vehicle, to for the oldest cars (notice the for the oldest cars).

Age of the policy

The claim severity and frequency age of the policy. The pure premium and premium. Note that this variable will not be used in premium models: most likely, younger policies (levels 0 – 4) would be a priori than older policies because of the severity. Moreover it only distinguishes five levels, with level 5 all policies that are at least 5 years old, so is not specified enough. This would be however a very useful variable for portfolio analysis of lossratio for example.
4.2 Univariate statistics and first insights

Table 4.3: Age of the policy, limited to 5.

<table>
<thead>
<tr>
<th>Age_policy</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5(+)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PctSum dur</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean y premium</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>costult</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>nclaim</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean n costult</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Age of the policyholder

For the age of the policyholder, values are observed in the range $0 - 112$; since the interval $[85, \ldots]$ represents only percent of the duration and the credibility of these levels is very low, the age is limited to 85. The weighted frequency histogram of age (figure 4.6) shows a local peak around the age of 30 and a global peak around 50 ($[48, 54]$ represents pct of the total duration).

Figure 4.6: Weighted frequency histogram of age of the policyholder.
4.2 Univariate statistics and first insights

**Figure 4.7:** Average claim frequency and CI for age.

**Figure 4.8:** Average claim severity and CI for age.
4.2 Univariate statistics and first insights

From the above figures, we can see what we already expected: the claims age, whereas also flat parts are observed for several intervals, for example. A trend is again observed from the age. A more discontinuous behaviour but also with an overall decreasing trend is observed for the severity; premium - older policyholders pay policyholders. Notice the wider confidence intervals for the smallest and largest values because of the lower total observed duration for these ages. Notice the for the pure premium in figure 4.9 so the frequency and severity combined, that is caused by the hidden drivers: the children of the policyholders that also drive the car, mostly around the age of 18 – 22.

As already mentioned, the given age classes in the variable age_class_cl are first of all not entirely correct and second, not so useful. Another segmentation that seems appropriate from the statistics here, is obtained in the next chapter.

4.2.3 Geographical zones

Details concerning the former segmentation Zone_cl can be found in appendix A.
Newzone

Table 4.4: Newzone, the extended geographical division.

<table>
<thead>
<tr>
<th>Zone</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>PctSum</td>
<td>dur</td>
<td>□ □ □ □ □ □ □ □ □</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

This division seems an improvement of the old considering the trend in claim frequency. This trend is almost linear, see the linear interpolation in figure 4.10. Clearly the premium follows these indications that zones are definitely more □ □. In figure 4.11 the influence of the sum of duration is visible: the intervals become larger for the zones with less duration.

Figure 4.10: Average claim frequency and 95% CI for newzone. A linear trend is added to fit the average claim frequency, this line has equation claim frequency = □ □ □ □ □ newzone +□ □ (computed with Excel).
4.2 Univariate statistics and first insights

Figure 4.11: Average claim severity and 95% CI for newzone.

Figure 4.12: Average pure premium, and CI, and retained premium for newzone; linear trend
pure premium = \newzone \times \text{newzone} + \text{.}

Notice that zones 4 – 9 sum up to only \% of the total duration but have a share of
\% in the total sum of premium and almost \% in the total sum of ultimate costs
(figure 4.13).
4.2 Univariate statistics and first insights

Figure 4.13: Share in total sum of premium (left) and ultimate costs (right) for different levels of newzone.

ZoneG

ZoneG, the social/economical/geographical division in 35 zones, contains a level 0, consisting of all observations where the zone is unknown (pct of the total duration), and a level U99, consisting of known but unclassified observations (pct). Other levels are always denoted by a letter-number combination where the letter is used several times (for example the first four levels are A01, A02, A03, A04). An explicit list describing the different levels can be found in Appendix A.

Often the grouped levels per letter, so A-H, have similar behaviour. From the figure we can see a more or less increasing trend in the mean number of claims.

Every level, apart from A01 and H-levels, contain at least a share of pct in the duration. B07 contains the most with almost pct; B and D represent total duration.
4.2 Univariate statistics and first insights

Figure 4.14: Average claim frequency and 95% CI with respect to zoneG.

Figure 4.15: Average claim severity and 95% CI with respect to zoneG (no costs were observed for U99 so this zone is not displayed).
4.2 Univariate statistics and first insights

Figure 4.16: Average pure premium, and CI, and retained premium with respect to zoneG (the observed premium for $U^{99}$ is $\square$).

Notice also the increasing width in for example figure 4.14 of the confidence intervals caused by the decreasing duration of the respective levels. The claim severity follows a more fluctual trend. Of the 35 levels, 8 levels represent about $\square$ pct of the total sum of premiums and ultimate costs (figure 4.17).

Figure 4.17: Share of different levels of zoneG in total sum of premium (left) and ultimate costs (right).
4.2.4 Vehicle characteristics

More detailed tables and values can be found in Appendix A.

**Code: sport, diesel and capacity**

Only \( \text{\%} \) \% of the duration is represented by **sport cars**, and those \( \text{\%} \) \% they have on average a \( \text{\%} \) \% claim frequency (\( \text{\%} \) \% versus \( \text{\%} \) \% for non-sport) although the claim severity is \( \text{\%} \) \% versus \( \text{\%} \) \% for non-sport). This variable will not be used in the premium models because of the very small group of sport cars.

Most vehicles in our dataset use **diesel** \( \text{\%} \) \% in comparison to petrol, where the diesel group premiums are \( \text{\%} \) \% on average. The diesel cars have a \( \text{\%} \) \% expected claim frequency (\( \text{\%} \) \% versus \( \text{\%} \) \% for petrol) and a mean pure premium of \( \text{\%} \) \% \% than for petrol. The claim severities \( \text{\%} \) \% \%. Diesel will be used for further model analysis.

For the **capacity**, \( \text{\%} \) \% of the data have registered 6, and \( \text{\%} \) \% can transport 5 persons. The other groups represent \( \text{\%} \) \% for 4 persons and less, and almost \( \text{\%} \) \% for 7 – 9 persons. No remarkable trends in pure premium, nor claim frequency or severity, nor premium can be seen. This variable will thus not be used for modelling.

**Power of the vehicle**

It seems that the categorization of **power** could be better since the average number of claims and cost ultimate does not show a trend over the different levels of pow_cat. To consider power as a continuous variable, we grouped all levels of 30\( \text{kW} \) or lower into one category, and the same for all values of 150\( \text{kW} \) or higher, because of the few observations in these extreme levels. However, we can see \( \text{\%} \) \% \( \text{\%} \) \%(figure 4.18). The retained premium shows \( \text{\%} \) \%. It turns out that almost \( \text{\%} \) \% of the duration is represented by vehicles with a power between 50 and 70\( \text{kW} \). All other levels represent each about \( \text{\%} \) \% \( \text{\%} \) \%(except for \([0, 40\text{kW}]\) with \( \text{\%} \) \% and \( [90, 100\text{kW}]\) with \( \text{\%} \) \%). The average claim frequency, the premium shows an \( \text{\%} \) \% for \( \text{\%} \) \% power, \( \text{\%} \) \% ultimate cost (see figure 4.20). Apparently a few decennia back, there was a significant trend visible between the claim frequency and the power of the vehicle, leading to a tariff that depended directly on the power. This was observed in the whole market, and nowadays the frequency trend is fully dissappeared but
however no insurance company dares to eliminate this factor from the tarification.

**Figure 4.18:** Average claim frequency (and CI) for the power of the vehicle, limited to the range of 30 to 150.

**Figure 4.19:** Average claim severity (and CI) for the power of the vehicle, limited to the range of 30 to 150.
4.2 Univariate statistics and first insights

Figure 4.20: Average pure premium (and CI) and retained premium for the power of the vehicle, limited to the range of 30 to 150.

4.2.5 Policy characteristics

Cover

The claims, premiums and costs only refer to civil responsibility; however, a percent of the duration here originates from contracts that are combined with an extended coverage (full or limited). One can derive that on average about a discount is given on the TPL-premium in full omnium policies. Policies that cover TPL only, produce average claim frequency and severity. When pricing insurance, it would be natural to make the base premium, here the TPL premium, independent of whether an additional coverage is taken; possible discounts would be given on the premium of the additional coverage. Moreover the differences between omnium and TPL are relatively small, so this variable will not be used for modelling. It may be interesting after the modelling, to analyse the loss ratio for different coverages.

<table>
<thead>
<tr>
<th>Cover</th>
<th>Other</th>
<th>Limited omnium</th>
<th>Full omnium</th>
<th>TPL</th>
</tr>
</thead>
<tbody>
<tr>
<td>PctSum</td>
<td>dur</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean y</td>
<td>premium</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>costult</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean n</td>
<td>nclaim</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>costult</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
4.2 Univariate statistics and first insights

Overtake

We distinguish the case where the policyholder can show his claim history by referring to his former insurer, and it is a direct overtake of the contract, or when it is a new contract (per cent of the time) and the insurer didn’t have any history. However the data are used for acceptance analysis but will not be used here for determining the premium.

Received discount

Of all the data, not less than per cent of the duration comes from policies where a reduction was given upon the premium. More specifically, per cent was given a 10 per cent discount or less, another per cent was given 15 per cent discount and another per cent was given 20 to 25 per cent discount. At last, profited from a discount of at least 30 per cent. To perform analysis, the budget was regrouped to discount, which has levels 0, 3 = [0, 3], 5 = [3, 5], ..., 40 = [30, 50] (see figure 4.21). There is visible, discount the client, so the discounts (as seen over the whole dataset off course). This is an excellent variable for analysing the current pricing model but will not be used to determine the new premium here. To give an idea, figure 4.21 shows the average claim frequency, severity, pure premium and current premium (no confidence interval limits are included).

!fig1discount.JPG

**Figure 4.21:** Mean claim frequency, severity, premium and pure premium for various levels of discount (intervals of budget).
4.2.6 History of the policyholder and other

Sexe of the policyholder and use of the vehicle

Of the total duration, the frequency of male policyholders is \( m \) \% pct. No differences are found in comparison to the female population, although it is well known that in an interaction term with age, it becomes a very significant predictor since young males drive less safe than young females and this difference decreases then with age. However this variable can’t be directly used in tarification by law so we will not use this any more.

Only \( m \) \% pct of the duration represents jobrelated transport. This small group private transport: the average number of claims is \( m \) (versus \( m \) for private transport), the claim severity is almost the same on average resulting in a pure premium is \( m \) pct \( m \) \% (versus \( m \)). Because \( m \), this variable will not be used in the tarification modelling.

Claimfree indicator

\( m \) of the observed duration here is from a policy that has been claimfree for the past 5 years. These \( m \) pure premium, \( m \) claim severity is \( m \). This indicator will turn out to be very influential (but as we may already remark, it is of course correlated with the BM).

| Table 4.6: Claimfree indicator of the five years preceding the policy. |
|--------------------------|---|---|
| Ind_claim_5year | 0 | 1 |
| PctSum | dur | \[ m \] | \[ m \] |
| Mean y premium | | \[ m \] | \[ m \] |
| Mean y costult | | \[ m \] | \[ m \] |
| Mean y nclaim | | \[ m \] | \[ m \] |
| Mean y costult | | \[ m \] | \[ m \] |
Number of policies of the policyholder

<table>
<thead>
<tr>
<th>Table 4.7: Number of policies by the policyholder.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Npol</td>
</tr>
<tr>
<td>PctSum</td>
</tr>
<tr>
<td>dur</td>
</tr>
<tr>
<td>Mean y</td>
</tr>
<tr>
<td>premium</td>
</tr>
<tr>
<td>costult</td>
</tr>
<tr>
<td>nclaim</td>
</tr>
<tr>
<td>Mean n</td>
</tr>
<tr>
<td>costult</td>
</tr>
</tbody>
</table>

Clearly the clients that [have on average a claim frequency; however the claim severity]. The total share in premium and ultimate costs is [for the four levels (figure 4.22)]. This variable can of course be used in the tarification.

Figure 4.22: Share in total sum of premium and ultimate cost for different levels of npol.

BM of the policyholder

The bonusmalus level of the policyholder is an interesting variable. A BM level of 0 is the most common with a frequency of [pct with respect to the duration. Levels 1 – 5 sum up to [pct, 6 – 11 up to [pct and only [pct has a level of 12 or higher (up to level 22). Note that [observed premium]
4.2 Univariate statistics and first insights

BM: the pure premium shows... Up to level 5, or even 11, the average number of claims... For the whole range this trend... which is easily explained by the fact that there are too little observations for higher levels and the attitude of the driver that, for a majority, does change (a bit) when paying such high premiums. This is visible in figure 4.23 by the wider confidence intervals for the claim frequency. The claim severity shows... over the levels;

Note that the BM will be used as a variable and not as offset, because we want to predict nclaim in a first step - the tariff follows later - and the BM as offset would explain more how the current tariff was obtained. Moreover the BM is not the real BM and we will see whether it would be better to use ind_claim_5year instead or not, when we are building the model in chapter five.

Figure 4.23: Mean claim frequency (and CI) for BM. A linear curve to fit the predictions (computed with Excel) is added with equation claim frequency = \text{claim frequency} = \star \text{BM} + \star.
4.2 Univariate statistics and first insights

Figure 4.24: Mean claim severity (and CI) for BM; linear curve with equation claim severity

\[ \text{claim severity} = \beta \times \text{BM} + \gamma. \]

Figure 4.25: Mean pure premium (and CI) and retained premium for BM.
4.3 Dependencies between the explanatory variables

When thinking about possible relations between the variables, we listed here a few remarks and insights. Note that this is purely informative concerning the relations between the explanatory variables themselves and does not necessarily say anything about the relation of these variables with the claim frequency nor severity. We didn’t list correlations since we aimed for insight in the variables rather then technical values. We used the SGPLOT (HBOX) statement of SAS that creates a horizontal box plot showing the distribution of a variable among the different levels of another variable, and takes into account weights (here the duration of course).

To illustrate its features, we take as example **age_car and discount / power**, see figure 4.26. The box itself represents the observations between the 25th (left border) and 75th (right border) percentile (so the length of the box is the interquartile range); the line in the box marks the median whereas the symbol denotes the mean. The lines that leave the box at the left (right) side stop at the minimum (maximum), or at a distance of 1.5 times the interquartile range of the box if the minimum (maximum) is smaller (bigger) than this value. Observations outside these *borders* are outliers and symbolised by the circles. As for the distribution of age_car, we see outliers for discount levels ; moreover a is visible to cars for discounts. Regarding the power, only deviations are noted for levels (i.e. power in and (i.e. power in where there are much fewer observed policies so this is not an important issue.
4.3 Dependencies between the explanatory variables

Figure 4.26: Boxplot of age_car (horizontal) for different levels of discount (left) and power (right). Note that a segmentation is used here for power that will be explained in the next chapter; to deliver a distribution analysis of age_car among the power, the original levels of power would be far too many.

These figures were made and analyzed for every pair of variables (so for $16 \times 15/2 = 120$ pairs, leaving ind_omnium and ind_reduction out since we look at cover and discount), but now we will only discuss couples of variables where a clear trend or shift was visible. Additional figures can be found in the appendix.

Dependencies with age_car

The relation with discount and power was already discussed above; as for the cover, there are many. For the full omnium, $\%$ pct of the policies concerns cars of $\text{years}$ years or younger; for the limited omnium this is $\text{years}$ years (figure 4.27). Civil responsibility has a distribution with a mean of $\text{years}$ years and $\%$ pct handles cars between $\text{years}$ years old. A distribution is observed for age_car (figure 4.27). In age_car is noticed for diesel cars: the median and mean is around $\text{years}$ for petrol, versus $\text{years}$ for diesel - the distribution is also more $\text{years}$ meaning that diesel cars cover a $\text{years}$ range of cars.
4.3 Dependencies between the explanatory variables

Figure 4.27: Boxplot of age,car (horizontal) for different categories of cover (left) and different levels of age (right).

Dependencies with age

With respect to ind_claim_5year, there is age of the policyholder. The claimfree population has a mean (and median) of about , with pct of the population being between ; population having had a claim in the past 5 years, with mean (and median) around , and pct between (figure 4.28). As for the number of policies, the monoclients (clients with npol equal to 1) do have a age on average but . The age of the policy for age, meaning that around the age of where the mean decreases from . is the distribution of age with respect to the BM (figure 4.28). Overall the expected is in fact observed: BM levels represent on average policyholders. The boxes, or 25 and 75 pct quartiles, are to each other and for most levels.
Dependencies between the explanatory variables

Figure 4.28: Boxplot of the age of the policyholder (horizontal) for ind_claim_5year (left) and boxplot of age (horizontal) for different levels of BM (right).

Dependencies with newzone

These figures are an excellent help of gaining insight in the categorisation of newzone and zoneG. The age of the car for newzone from an average age of around for zones 1 – 6 , to for zones 7 and 8 and for zone 9. Age_policy tells us that for zone, clients are found: is visible in the distribution (figure 4.29). Another factor that contributes to the fact that the zones should be categorised as risky: is observed, with an average age of in zone 9 to in zone 1 (figure 4.29). discount: the zones on average discount (on average pct for zone 9 to pct for zone 1). Finally, the relation between newzone and zoneG can be illustrated: we will elaborate about the levels of zoneG in the next subsection.
4.3 Dependencies between the explanatory variables

Dependencies with zoneG

First of all, it is clear that the zones are represented in the $E,F,G$ zones and definitely more in the $H$-zones (figure 4.30). Second, when taking a detailed look at all the levels of newzone for each zoneG, it turns out that zone is only represented in $H33$, zone only in $H31$, $H32$, $H33$ and zone almost only in $G$- and $H$- levels. Naturally from the link with newzone, npol behaves similarly for zoneG: are found in $B,C,D,E$- levels. As for the claimfree indicator, a of the average is observed (figure 4.30). As with newzone, the BM reveals , partly because the higher BM levels .

Figure 4.29: Boxplot of age,policy (left) and age (right) for the different levels of newzone (vertical axis).
4.3 Dependencies between the explanatory variables

Figure 4.30: Boxplot of ind_claim_5year (left) and newzone (right) for different categories of zoneG (vertical axis).

Dependencies with discount

As we already noted, discount is given to younger policies. There is also a visible discount for the BM, where the average discount decreases from for BM 0 to for BM 14. Naturally, the are given to younger policies; mostly then the clients were to a new contract or insurer. The number of policies that the policyholder closed with this insurer, has on the discount where the loyal clients (npol level 3/4) receive than the monoclients. In case the contract was an overtake, the discount is . Regarding the type of cover, an average pct is given to civil responsibility whereas the omnium discounts are (figure 4.31).
4.3 Dependencies between the explanatory variables

Figure 4.31: Boxplot of discount (horizontal) for different categories of cover (left) and different levels of overtake (right).

Dependencies with ind_claim_5year

The age of the policy is greater for the claimfree policies than for the non-claimfree policies over the last 5 years (on average versus ). is observed for npol (figure 4.32); note that these effects are age_policy / npol with the age of the policyholder. Whether the contract was an overtake or not claimfree indicator. The overtake policies represent only pct of the duration of these contract are not claimfree; non-overtake up to pct are not claimfree during the past 5 years.

Dependencies with BM

We already discussed the claimfree policies we find on average a BM than for the policies where a claim was reported in the last 5 years; this is . As for npol, a clients have a BM (figure 4.32). Note that the claimfree indicator and BM are naturally correlated since the BM is partly determined by the number of claims in the past 5 years. The BM is not always the correct, true BM as we already explained; for example for younger drivers this makes relations difficult to interpret since they may be claimfree (not having their license very long) and have a low BM (if their parents are good clients or some other reason) but may be intrinsically the most risky drivers.
4.3 Dependencies between the explanatory variables

Figure 4.32: Boxplot of npol (horizontal) for different categories of ind_claim_5year (left) and BM (right).

Other dependencies

The power of the vehicle, sport character and fuel type are related. Sport cars have a [power (figure 4.33)], naturally, and use petrol than diesel (because of the higher revolutions per minute). In this population the diesel cars have [than the petrol population, partly because the population of sport cars. The use of the vehicle [jobrelated transport seems to have a power on average (figure 4.33) and uses diesel than petrol (may be explained by the fact that diesel cars are more intended to drive longer distances with).
Figure 4.33: Boxplot of the power of the vehicle (horizontal) for different categories of code_sport (left) and private (right).
Chapter 5

Segmentation and interaction

In this chapter, we take a closer look at some variables and build our model assumptions. We start by dividing both the continuous variables and categorical variables with too many levels, in subgroups by choosing appropriate combinations. We can’t use of course all these original levels to appear in our further analysis - especially in interactions terms, too many levels make the model overparameterized (overfitted). Second, we will investigate the possible useful interactions: interfering variables that thus show different trends along different levels with respect to the response variable. The last section is an application of the Bühlmann-Straub estimator for the claim frequency regarding variable zoneG.

Note that from now on, we will always use weights in our model - as should be, the variance then increases for smaller weights. To study the variables in the first section, we will use the claim frequency model. We are aware that in order to segmentate for example age into categories, one would rather take into account the severity as well, as we aim for a pure premium estimate. But this is very difficult given the fact that we already start with many levels (for age there are  ); note that the frequency is however a very robust starting model given the large dataset. We will compare the AIC values to make justified choices (the smaller, the better).

As we already mentioned, we will not use all variables in the premium modelling process, see the next table for an overview. Only for the variables that will be used for modelling, will we take a detailed look at the possible interactions. Note that for example pow did not seem to influence the claim frequency or severity but may be interesting in an interaction with another variable, or does turn significant when other variables are already included in the model.
5.1 Variables and segmentation

The used strategy here is an iterative process where we look at the Poisson model, the computed parameter estimates, combined with the LSmeans statement in SAS, to decide which levels can be put together. Since age and BM are ordinal categorical variables, only consecutive levels can be grouped together. Several steps are necessary because on one hand, it is better to group only a few levels at the same time (2 or 3), and on the other hand, by regrouping the parameter estimates and lsmeans results change also so have to be reinterpreted as well.

The LSmeans statement computes Least Squares-means for any effect in a statistical model involving a categorical variable. With the diff statement, the differences between all means are computed, making a multiple comparison strategy possible. The standardized pairwise difference between level (or group) \( i \) and \( j \) of the considered variable, is defined by

\[
d_{ij} = \frac{\bar{y}_i - \bar{y}_j}{\hat{\sigma}_{ij}} \quad \text{with} \quad \hat{\sigma}_{ij} = s^2\left(\frac{1}{w_i} + \frac{1}{w_j}\right)
\]

with \( \bar{y}_i, \bar{y}_j \) the LSmeans for levels \( i, j \), \( s^2 \) the mean square error and \( w_i, w_j \) the sums of the weights of level \( i, j \). An approximate test is then constructed to test the null hypothesis that the associated population quantity, in our case the mean, equals zero - or the two groups having the same mean. This leads to a \( p \)-value: the probability of observing this test statistic or larger values under the assumed distribution (in case of a symmetric distribution it is larger in absolute value: \( Pr(Z > |d_{ij}|) \)).

### Table 5.1: Variables which will be possibly used and won’t be used for certain in the premium modelling process.

<table>
<thead>
<tr>
<th>Possibly in the model</th>
<th>Not in the model</th>
</tr>
</thead>
<tbody>
<tr>
<td>age_car</td>
<td>age_policy</td>
</tr>
<tr>
<td>age</td>
<td>code_sport</td>
</tr>
<tr>
<td>newzone</td>
<td>capacity</td>
</tr>
<tr>
<td>zonereg</td>
<td>cover</td>
</tr>
<tr>
<td>diesel</td>
<td>ind_omnium</td>
</tr>
<tr>
<td>pow</td>
<td>overtake</td>
</tr>
<tr>
<td>ind_claim_5year</td>
<td>discount</td>
</tr>
<tr>
<td>npol</td>
<td>ind_reduction</td>
</tr>
<tr>
<td>BM</td>
<td>private</td>
</tr>
</tbody>
</table>
Denote the initial model by model0, then we execute a few steps.

- Look at the parameter estimates for each level and compare consecutive levels: are the estimates close enough to each other? Then these levels make it to the shortlist of levels being possibly grouped together. Close enough means here in our personal intuition, close enough relative to the other estimates.

- Then look up the levels from the shortlist in the $p$-values of the lsmeans-output. They are only grouped together if the $p$-value is large enough - here normally larger than 10 pct, but the minimal attained $p$-value is stated in each step as Min $p$ in the tables.

- Take an additional look at the $p$-values, sorted the largest first for convenience, to check whether there are consecutive levels that score a very high $p$-value and were not already on the shortlist. These may also be grouped together.

After this process, the new variable is defined with less levels and model1 is created where this reduced variable is used. The same process can be repeated to create model2, and so on.

Note that this is certainly not a perfect and justified method; it is an intuitive method since it is important to reduce the number of levels before developing models or considering interactions. A few shortcomings we are aware of:

- Failing to reject the null hypothesis, so the $p$-value not being significantly small, does not mean that the alternative hypothesis is true, namely in this case that the group means are in fact equal. It only implies that the difference may not be large enough to be detected with the given sample size. We do however want to emphasize that this dataset is large so that differences should be detected (in theory - not all levels of age for example have a large enough sum of weights).

- In this way, the levels are only judged on its own predictive power with respect to the number of claims, while also the claim severity should be considered. Moreover, it is very plausible that the explanatory variable under consideration should or will appear in an interaction term later on. When doing this process for the interaction term, this may result in another segmentation.

5.1.1 Segmentation of the age of the policyholder

We performed these steps for the variable age, leading to the next table. Min $p$ is the minimal $p$-value that was used to obtain this segmentation in this step. Notice the decreasing
AIC values: clearly all the levels or too many levels is an overfit and not useful. After the second step, the AIC value increases but there was no other way: 16 levels are far too many to continue with, and the estimates showed not at all a linear trend (even not with a few corrections) so it was neither an option to try a continuous version.

Table 5.2: Consecutive steps (or segmentations) for age (Levels denotes the number of levels). The last line corresponds to the final, chosen segmentation as explained in the text.

<table>
<thead>
<tr>
<th>Step</th>
<th>Levels</th>
<th>AIC</th>
<th>Min p</th>
<th>grouped levels or intervals</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>69</td>
<td>-</td>
<td></td>
<td>every age $x$ is a level $[x]$</td>
</tr>
<tr>
<td>1</td>
<td>36</td>
<td>0.453</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>16</td>
<td>0.164</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>11</td>
<td>0.064</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>9</td>
<td>0.365</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>8</td>
<td>0.059</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5’</td>
<td>4</td>
<td>-</td>
<td></td>
<td>with a continuous variable and two dummy variables</td>
</tr>
</tbody>
</table>

After the fourth step, the AIC value increases again but here we found a better solution. We transformed the current segmentation into a continuous variable, meaning that this continuous age class has 9 levels, with values 1 to 9 corresponding to the 9 age classes as given in the table. Adding two dummy variables for the first two classes together and for the last one (so the binary variable equals 1 if the age is between and , and 0 otherwise; analogue for the binary variable ), this gives a model with only 4 parameters - an intercept, a slope and two parameters from the dummy variables. So actually the same number of classes is obtained in a different way, with less parameters to estimate and even a better AIC, see figure 5.1. (Note that a model with three dummy variables was also examined but did not give a lower AIC.)
Figure 5.1: Illustration of the small difference between the model from the fourth step (Est, L, U) with 9 levels (estimates) and the corrected model with 4 estimates - continuous with two dummy variables.

The final model is again illustrated in figure 5.2 where the empirical claim frequency, or model0, is displayed with their CI limits. The estimates from our final segmentation (the dotted line) lie in the CI. Note that we will use this final model and denote the age variable with age_cont, referring to the continuous part; but we will also use the purely categorical segmentation from the fourth segmentation, denoted with age_cat, as it will turn out to be significant in interactions with other variables (see the second section of this chapter).

Figure 5.2: Model0 with range from and its Est, L, U together with the final model - continuous with two dummy variables.
5.1.2 Segmentation of BM

Second, we also performed these steps for the variable BM, leading to the next table. Again the AIC values decrease; the marginal decrease gets smaller and smaller though. After the third step, two options are explored. A further reduction by one class gives a higher AIC, so a continuous model is also given here, containing a continuous BM class variable with range from 0 to 10 and corrections via dummy variables for BM equal to . So this model contains 8 parameters (slope and intercept and 6 from the dummy’s) - but still gives a higher AIC than the fourth segmentation step.

![Table 5.3: Consecutive steps (or segmentations) for BM.](image)

The problem here is that there are big differences for subsequent levels. For now, we will continue to work with the third segmentation model (11 levels). Recall that pct of the duration is represented by levels 0 – 5, so . Figure 5.3 shows the deviation from the observed claim frequencies for the original BM levels.
Figure 5.3: Model0 with □ levels and its Est, L, U together with the third segmentation model.

5.1.3 Segmentation of power

Then we performed the process for the variable pow, see the next table. Note that before we even began, we limited the range of power to $30 - 150$ as done in the previous chapter. Moreover all unpair values between 30 and 150 were coupled with the pair values below so the constructed intervals were of the form $[2n, 2n + 1]$ ($15 \leq n \leq 74$) - see the range on the $x$-axis in figure 5.4. Again the AIC values decrease up to a certain point.

As for age and BM, it seems that here again p-values below 0.062 lead to increasing AIC values. We will continue with the fourth segmentation, having 6 levels.
### Table 5.4: Consecutive steps (or segmentations) for pow.

<table>
<thead>
<tr>
<th>Step</th>
<th>Levels</th>
<th>AIC</th>
<th>Min p</th>
<th>grouped levels or intervals</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>61</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>1</td>
<td>24</td>
<td>0.257</td>
<td>0.120</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>13</td>
<td>0.120</td>
<td>0.240</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>8</td>
<td>0.240</td>
<td>0.156</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>6</td>
<td>0.156</td>
<td>0.059</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>0.059</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 5.4: The observed claim frequency for the 61 power levels from model0 with CI limits and estimates for segmentation step 4 with 6 levels.
5.1.4 Segmentation of age_car

The same steps are done for the variable age_car, see the next table. Note that before we even begin, we limited the range to 18 as explained in the previous chapter. Again the AIC values decrease up to a certain point. First, the p-value of 0.123 is apparently too low, finally, the last p-value of 0.048 also, meaning that the AIC value increased two times. We will continue with the third segmentation, having 5 levels, because this AIC value is even lower than the first segmentation.

<table>
<thead>
<tr>
<th>Step</th>
<th>Levels</th>
<th>AIC</th>
<th>Min p</th>
<th>grouped levels or intervals</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>19</td>
<td>-</td>
<td></td>
<td>levels</td>
</tr>
<tr>
<td>1</td>
<td>11</td>
<td>0.636</td>
<td>0.123</td>
<td>levels</td>
</tr>
<tr>
<td>2</td>
<td>7</td>
<td>0.123</td>
<td>0.269</td>
<td>levels</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>0.269</td>
<td>0.048</td>
<td>levels</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>0.048</td>
<td></td>
<td>levels</td>
</tr>
</tbody>
</table>

Table 5.5: Consecutive steps (or segmentations) for age_car.

Figure 5.5: The observed claim frequency for age_car, the 19 levels from model0 with CI limits and estimates for segmentation 3 with 5 levels.

Notice that this segmentation leads to a division in 5 levels with frequencies (in percentage of the total duration)
5.1.5 Linearity of newzone

As we already noted in the previous chapter, a linear trend can be observed in the claim frequency (and severity) with respect to newzone. When comparing the model with newzone as categorical variable (9 levels or degrees of freedom), and the model with newzone as continuous variable (2 degrees of freedom: intercept and slope), the second turns out the best. As can be seen in figure 5.6, there is a deviation for zones ... but a model with both the continuous newzone and these two dummy variables, is not better than the continuous term alone. So we conclude that newzone should be included as a continuous variable.

<table>
<thead>
<tr>
<th>Model</th>
<th>Degrees of freedom</th>
<th>AIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Categorical newzone</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>Continuous newzone</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Cont newzone and correction</td>
<td>4</td>
<td></td>
</tr>
</tbody>
</table>

Figure 5.6: Different models containing newzone: Est, L and U are obtained from the categorical model with 9 (original) levels. The continuous model (orange line) is equally good as the continuous model with two corrections (black dotted line).
5.2 Interactions of explanatory variables

Note that after the segmentation, age, BM, age_car, newzone and pow will denote the new variable from the chosen segmentation. An exception is made for age, we will consider both the continuous variable with two corrections (denoted with age_cont) and the categorical variable from the fourth segmentation (age_cat).

To check the interaction terms and possible significance, we perform several steps since it can not be done in a single, deterministic way. Moreover, it is not because a certain variable or combination of variables seems significant, that it truly is of significant importance to predict the response variable. There may be unexplained variation in the data that is actually explained by other variables which are not included or even known, and there may be a random variation that can not be explained by any variable. By ‘first order variables (estimates)’ we mean the (estimates of the) variables apart (variable1 variable2), whereas the ‘second order’ is then the interaction (variable1*variable2).

5.2.1 Two categorical variables

Types of interactions

Before proceeding, we take a closer look at the type of interactions that can appear here: two categorical variables, two continuous or one of each. Following table illustrates the different parameters for these three different interaction types: the upper horizontal row and left vertical column denotes the levels of the variables where capital letters denote the levels of a categorical variable, and numbers denote the levels of a continuous variable. Then the parameter estimates are given in the table, for every combinations of levels. For example if the levels of two categorical variables are (A,A’), the interaction estimate is $\alpha_1$, which is possibly different from other estimates $\alpha_2$ (A,B’), $\alpha_3$ (A,C’), and so on.
5.2 Interactions of explanatory variables

Table 5.7: Possible interaction types: two categorical variables (levels denoted by capital letters) with 9 estimates, a continuous and a categorical variable with 6 estimates, and two continuous (levels denoted by numbers) with 3 estimates ($\alpha$ for the interaction, $\beta$ and $\gamma$ for the variables alone). Note that there is always a parameter that can be left out because a certain level is chosen as reference level.

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>A'</td>
<td>$\alpha_1$</td>
<td>$\beta_1$</td>
<td>$\gamma_1$</td>
</tr>
<tr>
<td>B'</td>
<td>$\alpha_2$</td>
<td>$\beta_2$</td>
<td>$\gamma_2$</td>
</tr>
<tr>
<td>C'</td>
<td>$\alpha_3$</td>
<td>$\beta_3$</td>
<td>$\gamma_3$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$\alpha + \alpha_0$</td>
<td>$\beta + \beta_0$</td>
<td>$\gamma + \gamma_0$</td>
</tr>
<tr>
<td>2</td>
<td>$2\alpha + \alpha_0$</td>
<td>$2\beta + \beta_0$</td>
<td>$2\gamma + \gamma_0$</td>
</tr>
<tr>
<td>3</td>
<td>$3\alpha + \alpha_0$</td>
<td>$3\beta + \beta_0$</td>
<td>$3\gamma + \gamma_0$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>1’</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1’</td>
<td>$\alpha + \beta + \gamma$</td>
<td>$2\alpha + \beta + 2\gamma$</td>
<td>$3\alpha + \beta + 3\gamma$</td>
</tr>
<tr>
<td>2’</td>
<td>$2\alpha + 2\beta + \gamma$</td>
<td>$4\alpha + 2\beta + 2\gamma$</td>
<td>$6\alpha + 2\beta + 3\gamma$</td>
</tr>
<tr>
<td>3’</td>
<td>$3\alpha + 3\beta + \gamma$</td>
<td>$6\alpha + 3\beta + 2\gamma$</td>
<td>$9\alpha + 3\beta + 3\gamma$</td>
</tr>
</tbody>
</table>

As the table makes clear, with two categorical variables there is total freedom since every combination of two levels has its own estimate. When building a model with the interaction and the first order variables, the estimates and standard errors of the first order variables will all be zero since there is no variation left that is not covered by the interaction term. Note that for the estimates here the order seems to be important: when the interaction term is not the first included, the estimates for the other variables will be different from zero - however when adding the estimates to obtain the total relativity per level, this would give the same results. Moreover the goodness of fit criteria are identical for models with and without the first order terms.

For the interaction of a continuous and a categorical variable, it is more difficult. We assume that the estimates of subsequent levels of the continuous variable increase by the same amount (apart from the intercept), but this is not evaluated with another variable. So per level of this other categorical variable, another estimate reflects the amount of increase: $\alpha$ for level A, $\beta$ for level B, $\gamma$ for level C. If the continuous variable would need an extra factor $\xi$, so independent of the categorical variable, this could just be added to each existing factor $\alpha, \beta, \gamma$. This reflects in the model: when adding the continuous variable
Interactions of explanatory variables

To the interaction, its estimates are again zero and the goodness of fit criteria remain indifferent. However if the categorical variable needs an extra factor, independent of the continuous variable, this would have to be included separately (in the table denoted with subscript 0). As a consequence, the goodness of fit criteria improve (mostly) when adding the categorical variable next to the interaction. Whether the categorical variable on its own is significant in the model, is something else and dependent of the dataset of course.

As for two continuous variables, we assumed thus that the subsequent levels of each \( (x = 1 - 2 - 3 \text{ and } y = 1’ - 2’ - 3’) \) increase by the same factor \( (x\beta \text{ and } y\gamma) \), but we don’t know whether these effects together (estimate \( xy\alpha \)) will still be a linear relationship (probably not). So here we should include both first order variables \( \beta \) and \( \gamma \)-above \( \alpha \)-and see whether they are significant or not. Since age_cont is the only continuous variables that appears here in an interaction, the only types we have to deal with are the first and third type.

Including first order terms

So now we know that in two cases out of the three types of interaction, one may have to add the first order variable(s). Another argument that can be made, is a technical one. Many papers conclude that the first order should be added to preserve the invariance to location shifts. If predictors are centered by their mean, we would have [2]

\[
y_i = \beta_0 + \beta_i x_1^i + x_1^i \ast x_2^i + \epsilon_i \rightarrow \hat{x}_1^i \ast \hat{x}_2^i = (x_1^i - \bar{x}_1) \ast (x_2^i - \bar{x}_2) = x_1^i x_2^i - x_1^i \ast \bar{x}_2 - \bar{x}_1 \ast x_2^i + \bar{x}_1 \ast \bar{x}_2
\]

so the first order variables would indeed be included. However we work here with variables that may not be centered as that would be useless given the categorical character. We conclude that in case of two categorical variables, the variables alone will never be included since this doesn’t add any value; as for the other two types, we will see what the models decides for us and check the significance.

Results here

We perform a type3 test on the frequency (severity) model that contains two categorical variables, variable1 and variable2, and their interaction term variable1*variable2. The type3 test checks the significance: its \( p \)-values are determined by comparing the model of choice with the full model - meaning it includes all other variables specified, except the one in question. As opposite to the type1 test, which compares the reduced models with all variables included that are listed before the one in question, with and without the one
in question. So moreover the order does matter for the last type of test! This is only useful if there is a clear order of importance in the variables. Take for example the type3 test for variable1*variable2, then we always tested here the hypothesis $H_0: \beta_{\text{var1}\ast\text{var2}} = 0$ versus $H_A: \beta_{\text{var1}\ast\text{var2}} \neq 0$ by comparing the model that includes variable1, variable2 and variable1*variable2, with the model that includes variable1 and variable2. We already noted that terms may appear to be significant in a model but in reality are not in fact; we believe somehow in the inverse - if the interaction term is not significant in this limited model (only two distinct variables are used), then it is probably not significant in an extended model. When certain terms are found to be significant, we can not conclude that we found a causal relationship with the response variable, but at least we found indicators that it may be the case. Since we will combine a frequency and severity model, we have checked all the models both for the Poisson model with weight dur and the gammamodel with weight nclaim. The following table contains all combinations of pairs out of the 9 variables whose interaction turned out to be significant on a 5% level.

Table 5.8: Significant (on a 5% level) interaction terms in the frequency and severity model.

<table>
<thead>
<tr>
<th>Frequency model</th>
<th>Severity model</th>
</tr>
</thead>
<tbody>
<tr>
<td>age_car*ind_claim_5year</td>
<td>age_cat*ind_claim_5year</td>
</tr>
<tr>
<td>age_car*age_cat</td>
<td>age_cat*diesel</td>
</tr>
<tr>
<td>age_car*age_cont</td>
<td></td>
</tr>
<tr>
<td>BM*age_cont</td>
<td></td>
</tr>
<tr>
<td>diesel*pow</td>
<td></td>
</tr>
</tbody>
</table>

For example for the first interaction where we include age_car*ind_claim_5year, age_car, ind_claim_5year in the model, the estimates for age_car and ind_claim_5year are all zero as we explained before, but we have to include them to be able to perform the type3 test!

The next table lists the interactions where the type3 test could not be executed because the algorithm did not converge: the negative of Hessian matrix, which determines the variances and covariances of the estimators, is not positive definite. This means that the entries on the main diagonal are zero or negative - so the variance of some estimator(s) is zero or negative. Since the type3 test statistic of an estimator is computed with its standard error, it can not be calculated. Mostly (in this case it was) the standard errors of some estimators are zero - meaning there isn’t really any variation left in the data for this effect. This is the result of too many levels (overspecification), here caused by the interaction term, where there are no observations to cover this level, to calculate an estimate, confidence interval limits or standard error. Some statisticians on the SAS
Forum would say that the responsible effect or variable should be kept however, but here we choose not to. First of all, the type3 test couldn’t be performed so the significance is not confirmed, and second, many levels lead more to a significant result just because they have many levels - which is not always useful for a good tarification.

**Table 5.9:** Interaction terms where the algorithm did not converge because of too many levels containing few or none observations (only occured for the frequency model).

<table>
<thead>
<tr>
<th>Frequency model</th>
<th>age_cat*pow</th>
</tr>
</thead>
<tbody>
<tr>
<td>zoneG*age_car</td>
<td>age_cat*BM</td>
</tr>
<tr>
<td>zoneG*age_cat</td>
<td></td>
</tr>
<tr>
<td>zoneG*BM</td>
<td></td>
</tr>
<tr>
<td>zoneG*age_cont</td>
<td></td>
</tr>
<tr>
<td>zoneG*npol</td>
<td></td>
</tr>
</tbody>
</table>

It may come as a surprise that no more interactions turn out to be significant but often the variables on their own follow a trend that is not very different for various levels of another variable. We will illustrate this with ind_claim_5year and newzone. In figure 5.7 it is clear that the observed claim frequency of newzone follows a similar trend for both for ind_claim_5year equal to 0 or 1 as it does for the total population (dashed curve denoted with ‘newzone only’). Add the difference in claim frequency for claimfree and non-claimfree drivers (dashed curves denoted with ‘(not) claimfree only’) and it is clear that the interaction adds no extra value to the model.
5.2.2 Three categorical variables

In a second step (and last), we perform a type3 test on the frequency (severity) model that contains three categorical variables, variable1, variable2 and variable3. To make sure the significance is not due to lower degree combinations, the interactions variable1*variable2, variable1*variable3, variable2*variable3 are also included. The type3 test for the term variable1*variable2*variable3 delivers then the $p$-value that is determined by comparing the model including all the variables alone and their interactions, with the model including all these terms added with variable1*variable2*variable3. All combinations were tested, except for zoneG, and none of the third order terms turned out to be significant on a 5% level.

5.3 Application: BS estimators for zoneG

When modelling the claim frequency, our parameter $p$ in the Tweedie model equals 1 and $k$ denotes the number of levels, 35 for variable zoneG. The formulas as given in chapter three are calculated with SAS and Excel and result in

\[
\hat{\sigma}^2 = \text{[Formula]}
\]

\[
\hat{\tau}^2 = \text{[Formula]}
\]
So the within-group variance $\hat{\sigma}^2$ is greater than the between-group variance $\hat{\tau}^2$, indicating that the subpopulations determined by different levels, are different from each other. In figure 5.8 we presented the results along with the results from a Poisson model as given in the previous chapter: Est, L and U. The credibility estimator is the Bühlmann-Straub estimator as determined by the credibility factor (upper dashed line, calibrated to the right vertical axis), this is the weight given to the mean of this subpopulation; the estimated variance of each group ($\hat{\sigma}_i^2$) and population mean (constant line) are also displayed.

For the credibility factor, this is explained by the wider confidence intervals, indicating a smaller subpopulation (less data), and for also the large variation in this subpopulation. The same thing happens for the 1-levels and 2 and it is indeed easily verified that 3 and 4, along with the 1-levels and 2, have the same duration. Since the population the credibility estimators are almost equal to the observed subpopulation means (Est); only for 3 and 4 a difference can be seen and this is partly because subpopulations (Level variance), partly because of

Figure 5.8: Illustration of credibility theory for zoneG (the credibility factor is the only variable that refers to the right vertical axis).
Chapter 6

Building the model

We started in the previous chapter with an exploratory data analysis to gain insight on the behaviour of the dataset with respect to the number of claims and its subsets with respect to the explanatory variables. Now we will focus on predicting the number of claims or the claim frequency and try to use this to determine an optimal premium. First we will discuss the used method and assumptions, then we will build the frequency and severity model. To illustrate the difference, a pure premium model is also build. This is done in two different ways. The differences between those models mutually, and also with the original earned premium, are discussed in detail.

Recall that significance refers to and was always tested with the type3test as explained in the previous chapter in section 5.2.1. To illustrate the use of SAS, the subsequent steps of the first model, stepwise forward selection for the frequency, are added in SAS code.

6.1 Assumptions and preconditions

6.1.1 Explanatory variables

As already stated in the last chapter, we will only use 9 variables in our tarification model and the interactions that turned out to be significant (for frequency or severity); an intercept will be included. The goal is here to determine a model that leads to a tariff that is well segmentated but not too much, that is balanced with an appropriate number of different risk categories and hereby preserving the credibility of each level.
### 6.1 Assumptions and preconditions

<table>
<thead>
<tr>
<th>Table 6.1: Possible terms in the frequency and severity model.</th>
</tr>
</thead>
<tbody>
<tr>
<td>age_car</td>
</tr>
<tr>
<td>age_cont</td>
</tr>
<tr>
<td>age_cat</td>
</tr>
<tr>
<td>ind_claim_5year</td>
</tr>
<tr>
<td>diesel</td>
</tr>
<tr>
<td>pow</td>
</tr>
<tr>
<td>BM</td>
</tr>
<tr>
<td>newzone</td>
</tr>
<tr>
<td>npol</td>
</tr>
</tbody>
</table>

Note that the interaction terms include both age_cat and age_cont, as the other interaction (with the other age) was not significant. Recall that age_cont consists of the continuous term age_class and two corrections \[\text{and}\] . It turns out that for age_car*age_cont, the interactions with \[\text{and}\] are not significant while the variables alone are, so that for this interaction we would include age_car*age_class, age_car, \[\text{and}\] . (Recall that \(1 \leq \text{age_class} \leq 9\).)

#### 6.1.2 Method

As already mentioned we always use the pscale option so that the dispersion parameter \(\phi\) is estimated by Pearson’s estimator. A perfect model will never be obtained; we will make sure that age of the policyholder and some geographical zone will always be included. Mostly used is the \textbf{backward selection} process, where one intends to reduce the full model to a complete model, meaning a model with the best explanatory terms. So to begin, all possible variables are included in the model; then stepwise terms are excluded, every time the term which p-value is bigger than a certain significance level. The \textbf{forward selection} process does the opposite: it builds a model by extending an empty model, each step the variable is added that gives the best significance when included in the already composed model (always the p-value beneath the significance threshold of course). Another selection process is the \textbf{stepwise selection} where the same is done, backwards or forwards, but the variables are deleted or added based on their significance in their single factor model, and keeping them in the model if their significance in this model is below a certain, different from the other, threshold.

Backwards selection is not possible here because we have too many variables at our disposal: the model including all variables is not fully correct (negative of hessian matrix not
positive definite). We will both perform forward selection and stepwise forward selection. When the significance of a variable is maximal, meaning a p-value smaller than 0.0001, the AIC criterion will be minimized among the possible choices.

6.2 Frequency model selection

6.2.1 Stepwise forward selection

First we have to list all single factor models (models with only one variable or one interaction) using the variables or interactions that were significant in the frequency or severity model, to determine the order of adding the variables (or interactions) to our model. The threshold to remove existing model components can be set at 0.2 for example, but as it turns out this will not be important since the terms will remain significant at a 0.01% level.
### 6.2 Frequency model selection

<table>
<thead>
<tr>
<th>Variable in the frequency model</th>
<th>AIC</th>
<th>significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>BM<em>age_cont (i.e. BM</em>age</td>
<td></td>
<td>significant; BM, and age alone are not</td>
</tr>
<tr>
<td>BM<em>age_class BM</em>age</td>
<td></td>
<td>significant</td>
</tr>
<tr>
<td>BM</td>
<td></td>
<td>not significant (0.7)</td>
</tr>
<tr>
<td>age_cat*ind_claim_5year</td>
<td></td>
<td>significant</td>
</tr>
<tr>
<td>age_car*ind_claim_5year</td>
<td></td>
<td>significant</td>
</tr>
<tr>
<td>ind_claim_5year</td>
<td></td>
<td>significant</td>
</tr>
<tr>
<td>age_car*age_class age</td>
<td></td>
<td>significant; other interactions not</td>
</tr>
<tr>
<td>age_car*age_cat</td>
<td></td>
<td>significant</td>
</tr>
<tr>
<td>diesel*age_cat</td>
<td></td>
<td>not significant (0.1)</td>
</tr>
<tr>
<td>age_cont i.e.</td>
<td></td>
<td>significant</td>
</tr>
<tr>
<td>age_class age</td>
<td></td>
<td>significant</td>
</tr>
<tr>
<td>age_cat</td>
<td></td>
<td>significant</td>
</tr>
<tr>
<td>newzone</td>
<td></td>
<td>significant</td>
</tr>
<tr>
<td>zonereg</td>
<td></td>
<td>significant</td>
</tr>
<tr>
<td>npol</td>
<td></td>
<td>significant</td>
</tr>
<tr>
<td>diesel*pow</td>
<td></td>
<td>significant</td>
</tr>
<tr>
<td>age_car</td>
<td></td>
<td>significant</td>
</tr>
<tr>
<td>diesel</td>
<td></td>
<td>significant</td>
</tr>
<tr>
<td>pow</td>
<td></td>
<td>significant</td>
</tr>
<tr>
<td>Empty reference model</td>
<td></td>
<td>significant</td>
</tr>
</tbody>
</table>

Note that the two interactions that are not significant, age_cat*ind_claim_5year and diesel*age_cat, are included because they were significant for the severity model. When choosing to include every variable maximal one time, we thus have the following steps in building our model:
Table 6.3: Subsequent frequency models for stepwise forward selection, their AIC value and their significance (or p-value). The SAS code is included in the next subsection.

<table>
<thead>
<tr>
<th>Step</th>
<th>Variables in the frequency model</th>
<th>AIC</th>
<th>Significance of the variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>BM*age_class</td>
<td></td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td></td>
<td>BM*age</td>
<td></td>
<td>0.0017</td>
</tr>
<tr>
<td></td>
<td>BM*age</td>
<td></td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>1</td>
<td>BM*age_class</td>
<td></td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td></td>
<td>BM*age</td>
<td></td>
<td>0.0023</td>
</tr>
<tr>
<td></td>
<td>BM*age</td>
<td></td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td></td>
<td>age_car*ind_claim_5year</td>
<td></td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>2</td>
<td>BM*age_class</td>
<td></td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td></td>
<td>BM*age</td>
<td></td>
<td>0.0023</td>
</tr>
<tr>
<td></td>
<td>BM*age</td>
<td></td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td></td>
<td>age_car*ind_claim_5year</td>
<td></td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td></td>
<td>newzone</td>
<td></td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>3</td>
<td>BM*age_class</td>
<td></td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td></td>
<td>BM*age</td>
<td></td>
<td>0.0015</td>
</tr>
<tr>
<td></td>
<td>BM*age</td>
<td></td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td></td>
<td>age_car*ind_claim_5year</td>
<td></td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td></td>
<td>newzone</td>
<td></td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td></td>
<td>npol</td>
<td></td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>4</td>
<td>BM*age_class</td>
<td></td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td></td>
<td>BM*age</td>
<td></td>
<td>0.0004</td>
</tr>
<tr>
<td></td>
<td>BM*age</td>
<td></td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td></td>
<td>age_car*ind_claim_5year</td>
<td></td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td></td>
<td>newzone</td>
<td></td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td></td>
<td>npol</td>
<td></td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td></td>
<td>pow*diesel</td>
<td></td>
<td>&lt; 0.0001</td>
</tr>
</tbody>
</table>

So the fourth model includes actually all variables (possibly in interaction) - except for zoneG - leading to a model with $11 \times (9 + 2 + 2) \times (5 \times 2) \times 9 \times 4 \times (6 \times 2) = 617760$ different risk levels in theory, since there are 9 zones and 9 age levels. There are of course less parameters to estimate since not all the variables interact: including the intercept, scale parameter and excluding reference levels and zero estimates (for example for age equal to zero the estimates for BM*age are all zero), $1 + 1 + 11 \times 1 + (11 \times 2 - 1) + 11 +$
(5 \times 2 - 1) + 1 + (4 - 1) + (6 \times 2 - 1) = 69 \text{ parameters are calculated. When using these variables in the severity model, the variables are not all significant and the AIC score of the severity model is} \underline{\text{missing}}. 

When we would allow variables to occur multiple times, the first step would be the same but then \text{age car}^{\text{age class}} would be added instead of newzone. Note that only the interaction \text{age car}^{\text{age class}} is included, since the ‘other’ terms \text{age car} , \text{age}^{\text{age class}}, \text{age}^{\text{age}} would give estimates zero. This because they are already fully estimated in \text{age car}^{\text{ind\_claim\_5year}}, \text{BM}^{\text{age}}, \text{BM}^{\text{age}} which are already in the model. The other steps are again identical:
Table 6.4: Subsequent frequency models using stepwise selection when allowing variables to occur multiple times, their AIC value and their significance (or p-value).

<table>
<thead>
<tr>
<th>Step</th>
<th>Variables in the frequency model</th>
<th>AIC</th>
<th>significance of the variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>BM*age_class</td>
<td></td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td></td>
<td>BM*age</td>
<td>0.0017</td>
<td></td>
</tr>
<tr>
<td></td>
<td>BM*age</td>
<td></td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>1</td>
<td>BM*age_class</td>
<td></td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td></td>
<td>BM*age</td>
<td>0.0023</td>
<td></td>
</tr>
<tr>
<td></td>
<td>BM*age</td>
<td></td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td></td>
<td>age_car*ind_claim_5year</td>
<td></td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>2</td>
<td>BM*age_class</td>
<td>0.0040</td>
<td></td>
</tr>
<tr>
<td></td>
<td>BM*age</td>
<td>0.0069</td>
<td></td>
</tr>
<tr>
<td></td>
<td>BM*age</td>
<td></td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td></td>
<td>age_car*ind_claim_5year</td>
<td></td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td></td>
<td>age_car*age_class</td>
<td></td>
<td>&lt; 0.0036</td>
</tr>
<tr>
<td>3</td>
<td>BM*age_class</td>
<td>0.0009</td>
<td></td>
</tr>
<tr>
<td></td>
<td>BM*age</td>
<td>0.0067</td>
<td></td>
</tr>
<tr>
<td></td>
<td>BM*age</td>
<td></td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td></td>
<td>age_car*ind_claim_5year</td>
<td></td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td></td>
<td>age_car*age_class</td>
<td></td>
<td>&lt; 0.0065</td>
</tr>
<tr>
<td></td>
<td>newzone</td>
<td></td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>4</td>
<td>BM*age_class</td>
<td>0.0009</td>
<td></td>
</tr>
<tr>
<td></td>
<td>BM*age</td>
<td>0.0047</td>
<td></td>
</tr>
<tr>
<td></td>
<td>BM*age</td>
<td></td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td></td>
<td>age_car*ind_claim_5year</td>
<td></td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td></td>
<td>age_car*age_class</td>
<td></td>
<td>&lt; 0.0066</td>
</tr>
<tr>
<td></td>
<td>newzone</td>
<td></td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td></td>
<td>npol</td>
<td></td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>5</td>
<td>BM*age_class</td>
<td>0.0007</td>
<td></td>
</tr>
<tr>
<td></td>
<td>BM*age</td>
<td>0.0014</td>
<td></td>
</tr>
<tr>
<td></td>
<td>BM*age</td>
<td></td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td></td>
<td>age_car*ind_claim_5year</td>
<td></td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td></td>
<td>age_car*age_class</td>
<td></td>
<td>&lt; 0.0023</td>
</tr>
<tr>
<td></td>
<td>newzone</td>
<td></td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td></td>
<td>npol</td>
<td></td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td></td>
<td>pow*diesel</td>
<td></td>
<td>&lt; 0.0001</td>
</tr>
</tbody>
</table>
Note that the final result’s AIC value is not so much better as the previous model without the extra interaction term age_car*age_class. When using these variables in the severity model, the variables are not all significant and the AIC score of the severity model is [redacted]. Using a certain variable multiple times will not be used here for tariffication. First of all, the tariff is only more complicated using less available variables, so fewer risk categories are added but the computations do multiply. Second, clients with a bad value of this variable, get punished twice - this would not be easy to sell or explain.

Note that the data displays overdispersion: the variance is bigger than the mean and in theory, these should be equal for the Poisson distribution. This relation is thus rescaled by the parameter $\phi$, which is approximately [redacted] here.

### 6.2.2 SAS code: subsequent frequency models for stepwise forward selection

Different steps and final frequency model from stepwise model building:

```sas
PROC GENMOD DATA=DATA.DATATHESIS;
CLASS agecar4 (DESC ORDER=freq);
CLASS agew4 (DESC ORDER=freq) age (DESC ORDER=freq);  
CLASS age (DESC ORDER=freq) ind_claim_5year (DESC ORDER=freq); 
CLASS zoneg (DESC ORDER=freq);
CLASS pow6 (DESC ORDER=freq);
CLASS diesel (DESC ORDER=freq);
CLASS npol (DESC ORDER=freq);
CLASS bm4 (DESC ORDER=freq); 
WEIGHT dur;
MODEL nfreq=agew4 class*bm4 age*bm4 age*bm4 / D=POISSON PSCALE TYPE3 ;
ODS OUTPUT PARAMETERESTIMATES=PARMS;
ODS OUTPUT MODELFIT=MODELFIT ;
RUN;
```

```sas
PROC GENMOD DATA=DATA.DATATHESIS;
CLASS agecar4 (DESC ORDER=freq);
CLASS agew4 (DESC ORDER=freq) age (DESC ORDER=freq); 
CLASS age (DESC ORDER=freq) ind_claim_5year (DESC ORDER=freq); 
CLASS zoneg (DESC ORDER=freq);
```
6.2 Frequency model selection

CLASS pow6 (DESC ORDER=freq);
CLASS diesel (DESC ORDER=freq);
CLASS npol (DESC ORDER=freq);
CLASS bm4 (DESC ORDER=freq);
WEIGHT dur;
MODEL nfreq=agew4class*bm4 age**bm4 age**bm4 agecar4*ind_claim_5year
   / D=POISSON PSCALE TYPE3 ;
ODS OUTPUT PARAMETERESTIMATES=PARMS;
ODS OUTPUT MODELFIT=MODELFIT ;
RUN;

PROC GENMOD DATA=DATA.DATATHESIS;
CLASS agecar4 (DESC ORDER=freq);
CLASS agew4 (DESC ORDER=freq) age (DESC ORDER=freq);
CLASS age (DESC ORDER=freq) ind_claim_5year (DESC ORDER=freq);
CLASS zoneg (DESC ORDER=freq);
CLASS pow6 (DESC ORDER=freq);
CLASS diesel (DESC ORDER=freq);
CLASS npol (DESC ORDER=freq);
CLASS bm4 (DESC ORDER=freq);
WEIGHT dur;
MODEL nfreq=agew4class*bm4 age**bm4 age**bm4 agecar4*ind_claim_5year
   newzone / D=POISSON PSCALE TYPE3 ;
ODS OUTPUT PARAMETERESTIMATES=PARMS;
ODS OUTPUT MODELFIT=MODELFIT ;
RUN;

PROC GENMOD DATA=DATA.DATATHESIS;
CLASS agecar4 (DESC ORDER=freq);
CLASS agew4 (DESC ORDER=freq) age (DESC ORDER=freq);
CLASS age (DESC ORDER=freq) ind_claim_5year (DESC ORDER=freq);
CLASS zoneg (DESC ORDER=freq);
CLASS pow6 (DESC ORDER=freq);
CLASS diesel (DESC ORDER=freq);
CLASS npol (DESC ORDER=freq);
CLASS bm4 (DESC ORDER=freq);
6.2 Frequency model selection

\[ \text{WEIGHT dur;} \]
\[ \text{MODEL nfreq=agew4class*bm4 age**bm4 age***bm4 agecar4*ind_claim_5year} \]
\[ \text{newzone npol / D=POISSON PScale TYPE3 ;} \]
\[ \text{ODS OUTPUT PARAMETERESTIMATES=PARMS;} \]
\[ \text{ODS OUTPUT MODELFIT=MODELFIT ;} \]
\[ \text{RUN;} \]

\text{PROC GENMOD DATA=DATA.DATATHESIS;} \]
\[ \text{CLASS agecar4 (DESC ORDER=freq);} \]
\[ \text{CLASS agew4 (DESC ORDER=freq) age** (DESC ORDER=freq);} \]
\[ \text{CLASS age*** (DESC ORDER=freq) ind_claim_5year (DESC ORDER=freq);} \]
\[ \text{CLASS zoneg (DESC ORDER=freq);} \]
\[ \text{CLASS pow6 (DESC ORDER=freq);} \]
\[ \text{CLASS diesel (DESC ORDER=freq);} \]
\[ \text{CLASS npol (DESC ORDER=freq);} \]
\[ \text{CLASS bm4 (DESC ORDER=freq);} \]
\[ \text{WEIGHT dur;} \]
\[ \text{MODEL nfreq=agew4class*bm4 age**bm4 age***bm4 agecar4*ind_claim_5year} \]
\[ \text{newzone npol pow6*diesel / D=POISSON PScale TYPE3 ;} \]
\[ \text{ODS OUTPUT PARAMETERESTIMATES=PARMS;} \]
\[ \text{ODS OUTPUT MODELFIT=MODELFIT ;} \]
\[ \text{RUN;} \]

6.2.3 Forward selection

When doing forward selection, the resulting model is the same as with stepwise selection:
Table 6.5: Subsequent frequency models for forward selection (using each variable only once), their AIC value and their significance (or p-value).

<table>
<thead>
<tr>
<th>Step</th>
<th>Variables in the frequency model</th>
<th>AIC</th>
<th>significance of the variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>BM*age_class</td>
<td></td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td></td>
<td>BM*age</td>
<td>0.0017</td>
<td></td>
</tr>
<tr>
<td></td>
<td>BM*age</td>
<td></td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>1</td>
<td>BM*age_class</td>
<td></td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td></td>
<td>BM*age</td>
<td>0.0023</td>
<td></td>
</tr>
<tr>
<td></td>
<td>BM*age</td>
<td></td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td></td>
<td>age_car*ind_claim_5year</td>
<td></td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>2</td>
<td>BM*age_class</td>
<td></td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td></td>
<td>BM*age</td>
<td>0.0006</td>
<td></td>
</tr>
<tr>
<td></td>
<td>BM*age</td>
<td></td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td></td>
<td>age_car*ind_claim_5year</td>
<td></td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td></td>
<td>pow*diesel</td>
<td></td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>3</td>
<td>BM*age_class</td>
<td></td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td></td>
<td>BM*age</td>
<td>0.0005</td>
<td></td>
</tr>
<tr>
<td></td>
<td>BM*age</td>
<td></td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td></td>
<td>age_car*ind_claim_5year</td>
<td></td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td></td>
<td>pow*diesel</td>
<td></td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td></td>
<td>newzone</td>
<td></td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>4</td>
<td>BM*age_class</td>
<td></td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td></td>
<td>BM*age</td>
<td>0.0004</td>
<td></td>
</tr>
<tr>
<td></td>
<td>BM*age</td>
<td></td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td></td>
<td>age_car*ind_claim_5year</td>
<td></td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td></td>
<td>pow*diesel</td>
<td></td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td></td>
<td>newzone</td>
<td></td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td></td>
<td>npol</td>
<td></td>
<td>&lt; 0.0001</td>
</tr>
</tbody>
</table>

As explained before, here we only needed the performance of the single factor models in the first step to choose the first variable that was included: the interaction of age and BM had the lowest AIC score. Then every other variable (possibly interaction) that was listed in the table of single factor models, was added to this model and the significance was checked. If the added variable was significant, the AIC score was noted and so the (significant) variable with the lowest AIC score was included. This was repeated for each
step, requiring more computing time since every step the models are different (include one more variable) hence have to be reevaluated every time. It was possible that at a certain point, a certain variable was added, which was significant and performed the lowest AIC score, but made another variable, already in the model from a previous step, not longer significant on the chosen level. Then this no longer significant variable is excluded - this did not happen here for the frequency but it did for the severity (see further).

When using variables multiple times, the interaction age\_car*age\_class would also be included, again the same result as for stepwise forward selection.

### 6.3 Severity model selection

#### 6.3.1 Stepwise forward selection

Again we have to list all single factor models to determine the order of adding the variables (or interactions) to our model. The threshold to remove existing model components can be set at 0.2 for example.
### Table 6.6: Single factor severity models, their AIC value and their significance (or p-value), ordered with respect to AIC.

<table>
<thead>
<tr>
<th>Variable in the severity model</th>
<th>AIC</th>
<th>significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>zoneG</td>
<td></td>
<td>significant</td>
</tr>
<tr>
<td>diesel*age_cat</td>
<td></td>
<td>significant</td>
</tr>
<tr>
<td>BM*age_cont i.e.</td>
<td></td>
<td>only BM*age_class significant; others not</td>
</tr>
<tr>
<td>BM<em>age_class BM</em>age</td>
<td></td>
<td>significant</td>
</tr>
<tr>
<td>age_cat*ind_claim_5year</td>
<td></td>
<td>significant</td>
</tr>
<tr>
<td>age_car*age_cat</td>
<td></td>
<td>not significant (0.08)</td>
</tr>
<tr>
<td>newzone</td>
<td></td>
<td>significant</td>
</tr>
<tr>
<td>age_cat</td>
<td></td>
<td>significant</td>
</tr>
<tr>
<td>BM</td>
<td></td>
<td>not significant (0.1)</td>
</tr>
<tr>
<td>age_cont i.e.</td>
<td></td>
<td>only age_class significant; others not</td>
</tr>
<tr>
<td>age_class age</td>
<td></td>
<td>significant</td>
</tr>
<tr>
<td>age_car*age_class age car age</td>
<td></td>
<td>only age_class significant; others not</td>
</tr>
<tr>
<td>ind_claim_5year</td>
<td></td>
<td>significant</td>
</tr>
<tr>
<td>age_car*ind_claim_5year</td>
<td></td>
<td>not significant (0.2)</td>
</tr>
<tr>
<td>npol</td>
<td></td>
<td>significant</td>
</tr>
<tr>
<td>pow</td>
<td></td>
<td>not significant (0.3)</td>
</tr>
<tr>
<td>Empty reference model</td>
<td></td>
<td>not significant (0.1)</td>
</tr>
<tr>
<td>diesel</td>
<td></td>
<td>not significant (0.8)</td>
</tr>
<tr>
<td>age_car</td>
<td></td>
<td>not significant (0.7)</td>
</tr>
<tr>
<td>diesel*pow</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note that the interactions that are not significant, are included because they were significant for the frequency model; this is very clear for the last models since their AIC value is lower than the one of an empty model. When choosing to include every variable maximal one time, we build the model:
### Table 6.7: Subsequent severity models for stepwise forward selection (using each variable only once), their AIC value and their significance (or p-value).

<table>
<thead>
<tr>
<th>Step</th>
<th>Variables in the severity model</th>
<th>AIC</th>
<th>significance of the variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>zoneG</td>
<td>0.0005</td>
<td>0.0005</td>
</tr>
<tr>
<td>1</td>
<td>zoneG</td>
<td>0.0001</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td></td>
<td>age_cat*diesel</td>
<td>0.0001</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>2</td>
<td>zoneG</td>
<td>0.0001</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td></td>
<td>age_cat*diesel</td>
<td>0.0517</td>
<td>0.0706</td>
</tr>
<tr>
<td></td>
<td>npol</td>
<td>0.0469</td>
<td>so removed from the model</td>
</tr>
</tbody>
</table>

Note that age_cat scores better for the severity than age_cont; as we explained in the first section about segmentation of the previous chapter, the segmentation is only frequency-based. Would we have had a severity-based segmentation, probably age_cat would have been chosen over age_cont.

When allowing variables to occur more than once in our model, we end up with a model with an extra interaction term:

### Table 6.8: Subsequent severity models for stepwise forward selection (using variables multiple times), their AIC value and their significance (or p-value).

<table>
<thead>
<tr>
<th>Step</th>
<th>Variables in the severity model</th>
<th>AIC</th>
<th>significance of the variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>zoneG</td>
<td>0.0005</td>
<td>0.0005</td>
</tr>
<tr>
<td>1</td>
<td>zoneG</td>
<td>0.0001</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td></td>
<td>age_cat*diesel</td>
<td>0.0001</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>2</td>
<td>zoneG</td>
<td>0.0001</td>
<td>0.0517</td>
</tr>
<tr>
<td></td>
<td>age_cat*ind_claim_5year</td>
<td>0.0706</td>
<td>0.1110</td>
</tr>
<tr>
<td></td>
<td>age_cat*age_cat</td>
<td>0.7008</td>
<td>so removed from the model</td>
</tr>
<tr>
<td>3</td>
<td>zoneG</td>
<td>0.0001</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td></td>
<td>age_cat*diesel</td>
<td>0.0492</td>
<td>0.0765</td>
</tr>
<tr>
<td></td>
<td>npol</td>
<td>0.0528</td>
<td>so removed from the model</td>
</tr>
</tbody>
</table>
6.3.2 Forward selection

When performing forward selection, the same model is obtained with variables zoneG, age_cat*diesel and npol; age_cat*ind_claim_5year is not included because when joining the model, this interaction is not significant due to the fact that zoneG and age_cat*diesel are already included.

6.4 Tariff - Pure premium modelling

To study and compare tariff models, we will limit ourselves to the first 5 terms as resulting from the stepwise forward frequency model:

\[ \hat{\mu}_i = \beta_1(BM \times age_{class})_i + \beta_2(BM \times ag_{class})_i + \beta_3(BM \times ag_{class})_i + \beta_4(age_{car} \times ind_{claim \_5year})_i + \beta_5 newzone_i. \]

This leads to a model with \( 11 \times (9 + 2 + 2) \times (5 \times 2) \times 9 = 12870 \) different risk levels in theory, since there are 9 zones and 9 age levels. As we already noted, there are less parameters to estimate: including the intercept, scale parameter and excluding reference levels and zero estimates (for example for age equal to zero the estimates for BM*age are all zero), \( 1 + 1 + 11 \times 1 + (11 \times 2 - 1) + 11 + (5 \times 2 - 1) + 1 = 55 \) parameters are calculated.

6.4.1 Based on the frequency model

A tariff can directly be derived by multiplying a certain base premium with the predicted mean number of claims for the considered tariff cell; the base premium will have to be determined by a lossratio analysis of the total set of policyholders and by outlook of the company’s specific goals. This will be referred to as the frequency premium as it will be a tariff that doesn’t take into account the claim severity.

So then we would have, with \( \hat{\mu}_i \) the predicted mean frequency for observation \( i \) :

\[
\sum_i premium_i = base \ premium \sum_i \hat{\mu}_i = \frac{Expected \ Loss}{LR_{goal}} \Rightarrow base \ premium = \frac{Expected \ Loss}{LR_{goal}} \frac{1}{\sum_i \hat{\mu}_i}.
\]

When specifying the frequency model, SAS delivers the parameter estimates and predicted values. The predicted values then have to be multiplied with the duration for this policyholder to know the estimated frequency of this contract, to be able to make the
correct summation $\sum_i \hat{\mu}_i$. Using the same LR as obtained with the original premiums of this dataset, we find the base premium as:

$$\text{base premium} = \frac{1}{\sum \hat{\mu}_i} = \ldots$$

This can then be multiplied with the frequency estimate of the reference level (determined by the most observed level of a variable) to obtain the base tariff. The reference level is here a policyholder living in geographical zone 1, between years old, driving a car of years, with a claim in the past 5 years. This reference level has $\hat{\mu} = \ldots$ so that the base tariff would be

$$\text{base tariff} = \text{base premium} \cdot \hat{\mu}_{\text{ref.level}} = \ldots$$

This tariff then has to be multiplied with the according factors of the explanatory variables. Note that the used $\hat{\mu}_{\text{ref.level}}$ is than the average claim frequency; so we expect most factors to be This does not matter in fact for the tariff; the base tariff is not an absolute minimum or average tariff; it is the tariff of the reference level which is chosen by the model specification options.

First there is the factor of the age in interaction with the BM level - we may not forget the difference between the categorical and continuous variables. For example for the class, which is actually of the continuous variable age, one would have

$$\ln (\hat{\mu}) = \sum \beta_{\text{ref.level}} + 5 \cdot (\ldots) = \sum \beta_{\text{ref.level}} + \ldots$$

$$\iff \hat{\mu} = \exp \left( \sum \beta_{\text{ref.level}} + 5 \cdot (\ldots) \right) = \hat{\mu}_{\text{ref.level}} \cdot 5 = \hat{\mu}_{\text{ref.level}}$$

since the estimate is the amount by wich $\ln(\hat{\mu})$ will increase for each increase of age by one unit.

Attention for age, or the age intervals and, where the interactions age*B and age*B add an extra estimate or factor. The next tables will always give the factor for $\hat{\mu}$ (so after taking the exponential).
Table 6.9: Factors in the (frequency-based) tarification for the interaction age\_class*BM , age\_sex*BM and age\_sex*BM. Underlined factors are either bigger than \(1\), or smaller than \(1\).

<table>
<thead>
<tr>
<th>age/BM</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Then there is the factor of age\_car\_ind\_claim\_5year:

Table 6.10: Factors in the (frequency-based) tarification for the interaction age\_car\_ind\_claim\_5year.

<table>
<thead>
<tr>
<th>ind_claim_5year/age_car</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

And finally the geographical zone has to be included:

Table 6.11: Factor in the (frequency-based) tarification for newzone (continuous, so the factors are powers of the estimate \(1\)).

<table>
<thead>
<tr>
<th>newzone</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

When comparing the difference between the earned premium (as paid by the policyholder, so after possible discounts) and this frequency premium for the different variables
or subpopulations, we look at the ratio

\[
diff_i = \frac{\sum G_i \text{ premium} - \sum G_i \text{ premium}_{model}}{\sum G_i \text{ premium}}.
\]  

(6.1)

Here \( G_i \) stands for certain subpopulations; see further in section 6.4.3.

Most observations are always located in the interval \( \) . For age, there are some clear trends which are shown in the following figure 6.1: drivers do . The oldest drivers (equal to ) however seem to pay (middle figure) and this is by age_class (since they have ). The second figure 6.2 indicates that the BM levels have difference in premium whereas have the . As for age_car, the experience , in opposite to the year old cars; this was already expected from the difference in estimated coefficients.

freqpremage.JPG

**Figure 6.1:** The relative difference of the modelled frequency premium with the earned premium. From left to right: age, age, age_class.
For the claimfree indicator a small shift is observed for level 1. Most observations are in $[\boxed{0.1}, \boxed{1}]$, for level 0 $\boxed{1}$, the observations are still relatively dense between $\boxed{0.1}$ and $\boxed{1}$. This was expected because of the estimated coefficients for this level (see tariff table of age_car*ind_claim_5year above).

### 6.4.2 Based on the frequency and severity model

When modelling the claim severity with a gamma model, using the same variables as the frequency model, the parameter estimates can be added (before taking the exponential, multiplied after) to obtain a model that predicts the pure premium - another way to determine a tariff. The predicted pure premium has to be multiplied again with the duration to know the correct estimate of the considered contract. This will be referred to as the \textit{freq-sev premium}.

However in this case, the LR is not preserved since the sum of these estimated pure premiums is a factor smaller than the original earned premiums. So the estimated pure premium is multiplied with this factor in order to obtain the same LR. Here the base tariff is then given by the estimated mean pure premium of the reference level, which is the same reference level as for the frequency model in the previous section, multiplied with the computed factor:

$$\text{base tariff} = \boxed{\text{base pure premium}} = \boxed{\text{factor}} = \boxed{\text{result}}$$
Note that this is a much higher base tariff than obtained for the frequency model so we would expect lower values in the factor tables of the different explanatory variables.

**Table 6.12:** Factors in the freq-sev tarification for the interaction age_class*BM , age*BM and age*BM.

<table>
<thead>
<tr>
<th>age/BM</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Then there is the factor of age_car*ind_claim_5year:

**Table 6.13:** Factors in the freq-sev tarification for the interaction age_car*ind_claim_5year.

<table>
<thead>
<tr>
<th>ind_claim_5year/age_car</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

And finally the geographical zone:

**Table 6.14:** Factor in the freq-sev tarification for newzone (continuous, so the factors are powers of the estimate).

<table>
<thead>
<tr>
<th>newzone</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Very similar trends are observed for the difference (ratio [6.4]) of the freq-sev premium with the earned premium, with respect to the levels of newzone, age, age, age_cont
and age\textsubscript{car} as with the frequency premium. Figure 6.3 shows the relative differences that display other trends. For \texttt{ind\_claim\_5year}, the premiums for level 1 relative differences than the frequency premium. The BM levels also show a couple of differences with the frequency premium in figure 6.2: for levels and the is while for levels the is. For level the is also slightly; these differences are explained by the premium factors as given in the tables above. More extreme values were underlined (smaller than , bigger than - in practice these kinds of difference would never be applied, certainly not to one factor (here BM*age). These factors will be furthermore referred to when comparing tariffs and discussing tarification in practice.

\texttt{freqsevprembm.JPG}

**Figure 6.3:** The relative difference of the modelled freq-sev premium with the earned premium. Left: \texttt{ind\_claim\_5year}, right: BM.

### 6.4.3 Comparison of the two models mutually, and with the earned premium

Now the dataset was divided in all possible groups forming different combinations of the explanatory variables BM, age\textsubscript{car}, age\textsubscript{class}, age\textsubscript{car}, ind\_claim\_5year and new\textsubscript{zone}. Note that this leads in this dataset to 4346 groups. These are numbered from 1 to 4346 and this group number is always on the horizontal axis in the figures, like figure 6.4. The groups are in this order of variables, with values smaller to larger. This total number of groups, 4346, is a lot less than the theoretical number of 12870 groups, meaning that not every possible combination is observed in this dataset. For each subgroup \(G_i\) we will
define several relative differences to illustrate the ratios of the premiums, mutually and with the earned premium as was paid by the policyholder to the insurer.

First we will look in figure 6.4 at the difference between the frequency premium and the freq-sev premium through the ratio (defined for each subgroup $G_i$)

$$diff_i = \frac{\sum_{G_i} frequencypremium - \sum_{G_i} freq-sev premium}{\sum_{G_i} frequencypremium}.$$

Figure 6.4: The relative difference of the modelled frequency premium with the freq-sev premium.

Most observations are between but for some groups the difference are larger or smaller. Note that the explanation is always that the combination of the factors (as given in the tables above) with the base premium gives a larger or smaller estimate. Frequency premiums are more than for the following groups.

- $G_{2050}$ is the group of policyholders with BM $\hat{1}$ and older than $\hat{1}$, as already seen, the factor BM*age influences strongly their estimated premium. For example for level $\hat{1}$ we would have for the frequency premium circa $\hat{1} \times \hat{1} = \hat{1}$, while for the freq-sev premium we would estimate $\hat{1} \times \hat{1} = \hat{1}$, which is of the frequency premium.

- $G_{3850}$ consists again of the older policyholders (age.cat with BM level $\hat{1}$).
6.4 Tariff - Pure premium modelling

- And $G_{3400}$ has contracts with BM and age cat equal to.

And frequency premiums are more than for the following groups.

- $G_{1700}$ which is BM level and age.

- $G_{2900}$ are again the oldest age category with BM - an extreme outcome was expected as the freq-sev estimate is.

- The groups $G_{3500} - G_{3600}$ contains the three oldest categories (so age with BM level . Again recall the very large freq-sev estimate of for.

- $G_{4100} - G_{4300}$ contains BM level for age and BM level for age . Again these are categories where estimates were calculated.

Second, we calculate the relative difference for each subgroup $G_i$ of the original earned premium and the modelled (frequency or freq-sev) premium:

$$diff_i = \frac{\sum G_i \text{ premium} - \sum G_i \text{ premium}_\text{model}}{\sum G_i \text{ premium}}.$$  

In figure 6.5 this relative difference is shown in a scatter plot over the 4346 groups, both for the frequency premium (blue) and the frequency-severity premium (green).
Figure 6.5: The relative difference of the modelled premiums with the earned premiums (blue for the frequency premium and green for the frequency-severity premium, values referring to the left vertical axis). These values are $\text{diff}_i$ as defined above and can be read on the left $y$-axis with range $[\,\,\,]$. The share in total duration is also given (red, values referring to the right vertical axis).

Clearly very similar trends are observed for both premiums in comparison to the earned premium. Recall that the total sum of premiums is the same for the two premiums as for the original premium.

Modelled premiums which are at least $\hat{\text{diff}}_i$ than the paid premiums are observed for the observations around $G_{2600}$, $G_{3400}$, $G_{3700}$ and $G_{3900}$ (figure 6.5).

- The first group $G_{2600}$ are policyholders with BM level $\text{BM}$ with age $\text{age}$. Probably due to their $\text{age}$, the asked premiums are $\hat{\text{diff}}_i$ than the predicted.

- $G_{3400}$ consists of policyholders with BM level $\text{BM}$ age_class $\text{age}$ in the interval $\text{age}$. Apparently these contracts are asked higher premiums than modelled, probably given the high BM.

- The third group $G_{3700}$ again is a group contracts with BM level $\text{BM}$ but now the $\text{policyholders}$, with age $\text{age}$.

- Finally $G_{3900}$ is the group where the policyholder has BM level $\text{BM}$. 
The fact that BM levels have here predicted premiums, is explained by the fact that there were observed for the claim frequency for these levels (see figure 4.23 in chapter four) and the earned premium and (figure 4.25).

Modelled premiums are for some groups than the paid premiums for the observations around \( G_{600} \), \( G_{2500} \), \( G_{2800} \) and \( G_{3600} \), and lots of observations between \( G_{1500} - G_{2100} \) and \( G_{3100} - G_{3300} \) (figure 6.5). These trends are more spread and not easy to divide in groups with certain characteristics. The claimfree indicator is mostly equal to for the premiums, also BM levels and policyholders seem to be premiums (see also the tariff factor tables above).

### 6.4.4 Tweedie modelling

We already discussed that the total loss can be written as \( \sum_{i=1}^{N} \), with \( N \) the number of claims and \( Z_i \) the severity of claim \( i \), namely a collective risk model; in this case a compound Poisson process as \( N \) is Poisson distributed (and \( Z_i \) gamma distributed). This model is a member of the Tweedie family, with parameter \( 1 < p < 2 \), and of the exponential family - meaning that it can be used in GLM. SAS provides a specification to identify this Tweedie distribution; however in our version this was not available and thus had to be programmed directly. This can be done with the deviance, as it is fully determined by the observed value \( y_i \), the weight \( w_i \), the mean \( \mu_i \) and the parameter \( p \).

To determine \( p \), we recall the parameter relations from the second chapter (2.11):

\[
\mu = \lambda \frac{\alpha}{\beta}, \quad p = \frac{\alpha + 2}{\alpha + 1}, \quad \phi = \frac{\lambda^{1-p}}{2 - p} \left( \frac{\alpha}{\beta} \right)^{2-p}.
\]

Clearly \( p \) will be between 1 and 2 (it is a decreasing function of \( \alpha \) that starts at 2 for \( \alpha = 0 \) and goes to 1 for \( \alpha \to \infty \)) and \( p \) depends only on \( \alpha \) thus only on the claim severity. We saw in chapter one that the coefficient of variation for a gamma distribution is \( 1/\sqrt{\alpha} \), so the closer the losses are clustered to the mean, the larger \( \alpha \) is and the closer \( p \) will be to one. Also the variance of the gamma distribution was \( \mu^2/\alpha \) or stated with the classical variance relationship \( \phi \mu^2 \), so \( \alpha = 1/\phi \). Thus we have

\[
p = \frac{1/\phi + 2}{1/\phi + 1} = \frac{\alpha + 2}{\alpha + 1}.
\]
We calculated for costult in our dataset (for the analysis weighted with nclaim), that \( \alpha = E^2/Var = \ldots \). So this results in a scale parameter \( \phi \) of \( \ldots \), or \( p = \ldots \). Note that here was assumed that \( \phi \) is constant where it actually varies with \( \lambda, \alpha \) and \( \beta \) as can be seen from (6.2). As explained in [11], \( \phi \) appears only in the variance and not in the expression for the mean (6.2), so that the estimator itself is not directly affected but the efficiency of the estimator is. For the large number of observations here, and the small range of \( \mu \), a constant \( \phi \) would be a good approximation. But we will do both here, a constant \( \phi \) and variable \( \phi \) approach.

The idea of programming the model based on the deviance in the GENMOD statement, was found in [13]. When including a deviance statement, the variance also has to be specified. The SAS code can be found in Appendix B, we programmed the functions

\[
Var = \frac{\phi}{w} \mu^p \quad \text{and} \quad Dev = \frac{w}{(1-p)(2-p)} \left[ y^{2-p} - y^{1-p}(2-p) + \mu^{2-p}(1-p) \right] \quad (6.3)
\]

where \( y \) and \( \mu \) are defined, for each observation, as the response and the predicted mean of this observation (\( w \) is the duration, as weight). Then the variable pure_premium (=costult/dur) is estimated with the explanatory variables BM*age_cont, BM*age, BM*age, age_car*ind_claim_5year and newzone.

**Constant \( \phi \)**

When using a fixed value for \( \phi \), we already calculated the \( \alpha \) parameter of the gamma distribution of costult above. However when performing some tests, we found that \( p \) values give a 

AIC score so we optimized \( p \) to obtain a (second) estimate. This gives rise to another \( \phi \) of course (value \( \ldots \)). With those two couples \( (p,\phi) = (\ldots,\ldots) \) and \( (\ldots,\ldots) \), the variance and deviance were defined through (6.3) and two models were obtained with the corresponding estimated pure premiums. For further reference, these will be called the ‘tweedie models’.

**Varying \( \phi \)**

For these two values of \( p \), we also performed the tweedie model with variable \( \phi \), so \( \phi \) was defined with the relation (6.2). Note that the dataset already contains the variables

- ‘freqpred’ : estimated frequency from the frequency model with the variables as stated above,
- ‘sevpred’ : the estimated severity from the severity model with those variables,
so that ‘freqpred’ becomes \( \lambda \) (the mean of the Poisson distribution) and ‘sevpred’ becomes \( \frac{\alpha}{\beta} \) (the mean of the gamma distribution) in the definition

\[
\phi = \frac{\lambda^{1-p}}{2-p} \left( \frac{\alpha}{\beta} \right)^{2-p}
\]

\[
\iff \phi_i = \frac{(freqpred_i)^{1-p}}{2-p} (sevpred_i)^{2-p}
\]

These models give of course rise to another set of predicted pure premiums and will be referred to as ‘tweedievarmodels’.

**Comparison of these two types of Tweedie models**

In the same way as done above for comparing the frequency and frequency-severity premium, we will graphically show the differences, for the subgroups defined by different combinations of the set of explanatory variables. In figure 6.6 we plotted

\[
diff_i = \frac{\sum_{G_i} \text{premium model}(p = ) - \sum_{G_i} \text{premium model}(p = )}{\sum_{G_i} \text{premium model}(p = )}.
\]

We can see that the difference between the two \( p \)-values for the tweediemodels is much than for the tweedievarmodels. In the subpopulations around \( G_{4000} \) a couple of groups are observed with outlier values. For the tweediemodel it concerns the groups with BM and age_cat levels ( difference) and For the tweedievarmodel it concerns the groups with BM and again age_cat levels and This corresponds to the difference in estimated parameters which is significantly for those levels.
Figure 6.6: The relative difference between the two $p$ values of the modelled premiums of the tweediemodels (with constant $\phi$, in blue) and of the tweedievarmodels (variable $\phi$, in red).

Relatively the difference between the $\phi$ values is also $\text{diff}_i$ for the tweediemodels. For fixed $\phi$, this is of course fixed, for variable $\phi$, we took the maximal difference over the groups $G_i$ and the difference of the summation over all $\phi$-values:

$$\text{diff}_i = \frac{\max_i G_i \phi(p = \text{var}) - \sum G_i \phi(p = \text{var})}{\sum G_i \phi(p = \text{var}) - \sum G_i \phi(p = \text{const})}.$$ 

Note that $\phi(p = \text{var})$ is always smaller than $\phi(p = \text{const})$; the minimal relative difference over the groups $G_i$ is $\text{diff}_i$.

From figure 6.7 we can see that the difference between the constant $\phi$ and variable $\phi$ is very similar for the two different $p$-values, defined as:

$$diff_i = \frac{\sum G_i \text{premium var}\phi - \sum G_i \text{premium const}\phi}{\sum G_i \text{premium var}\phi}.$$
6.4 Tariff - Pure premium modelling

Figure 6.7: The relative difference of the modelled premiums with constant $\phi$ and variable $\phi$ for $p$ equal to (blue) and (red).

For completeness, we give some goodness of fit statistics here for the different Tweedie models:

| Table 6.15: Goodness of fit statistics for the four different Tweedie models. |
|------------------|------------------|------------------|------------------|
| $p$              | Deviance          | loglikelihood    | AIC              |
| Tweediemodels    |                  |                  |                  |
| Tweedievarmodels|                  |                  |                  |

6.4.5 Comparison with earned premium

Now we can compare the estimated premiums from the tweedie(var) models with the earned premiums. Because the difference between the $p$-values is so small, we will continue with only one $p$-value, $p = \text{[value]}$. (Note that we did perform the following analysis also for $p = \text{[value]}$, and this confirmed that they were almost identical.) So in figure 6.8 the ratio

$$diff_i = \frac{\sum_{G_i} \text{premium} - \sum_{G_i} \text{premium}_{\text{tweediemodel}}}{\sum_{G_i} \text{premium}}.$$  

(6.4) is displayed. It is clear that the majority is between $\text{[value]}\%$. It turns out that the results are very similar to the difference between the earned premiums and the freq(sev)
6.5 Comparison of the frequency(severity) premium and the Tweedie premium

As we already performed several times, the modelled premiums are multiplied with a certain factor so that the LR would be identical to the LR of the dataset, if the modelled premiums were the actual premiums. For the tweedie premiums this factor is about \( \frac{\sum G_i \text{frequency premium} - \sum G_i \text{premium model}(p = \mathbf{\mu})}{\sum G_i \text{frequency premium}} \) is for the tweedie models and tweedievar models almost identical. Again this figure was also made for the other \( p \)-value ( \( \mathbf{\mu} \) ), but the results were identical (no difference would be visible on the figure).
6.5 Comparison of the frequency(severity) premium and the Tweedie premium

Figure 6.9: The relative difference of the frequency premiums and the tweedie(var) premiums for $p$ equal to $G$: the tweediemodel in blue (almost not visible) and tweedievarmodel in red.

Note that only for the observations around $G_{4150} - G_{4300}$ the trend is different for $\phi$ variable or constant. This concerns the subpopulations with BM level $\Phi$ and age between $G_{4150}$. Most observations are between $\Phi_{4150} - G_{4300}$, but there are lots of subpopulations outside this scope. The frequency premium is at least $G_{4150}$ than the tweedie premium for the following groups.

- $G_{1100}$ is the group with BM level $\Phi$ and age either $\Phi_{4150} - G_{4300}$, or $\Phi_{4150} - G_{4300}$ (so age_cat $\Phi_{4150} - G_{4300}$)
- $G_{2050}$ contains the $\Phi$ policyholders ($\Phi$) with BM level $\Phi$.
- $G_{3400}$ is the subpopulation with BM $\Phi$ and age_cat equal to $\Phi$ (so $\Phi$).
- Around $G_{4000}$ we find the $\Phi$ policyholders ($\Phi$) with BM $\Phi$ and the policyholders $\Phi$ with BM $\Phi$.
- And $G_{4300}$ contains BM $\Phi$ and the $\Phi$ and $\Phi$ policyholders.

But also, the frequency premium is at least $G_{4150}$ than the tweedie premium.

- $G_{600}$ is the group of $\Phi$ policyholders (age $\Phi$) with BM level $\Phi$
- $G_{1700}$ consists of age $\Phi$ and BM $\Phi$. 
6.5 Comparison of the frequency(severity) premium and the Tweedie premium

- Around $G_{2850}$ we find policyholders between [ ] years old with BM [ ].
- $G_{3500} - G_{3600}$ contains BM level [ ] with again age range [ ].
- The first of the three extreme subpopulations below in the figure, is $G_{2900}$, consisting of BM [ ] and the [ ] policyholders ( )
- Then $G_{3600}$ follows with again the [ ] policyholders, now with BM [ ]
- $G_{4200}$ is the third extreme group with BM [ ] and the [ ] policyholders [ ].

Now we illustrate the difference with the freq-sev premium in figure 6.10, we expect of course that this would be smaller since both premiums now take into account the claim severity (which was not the case for the frequency premium). So the difference

$$diff_i = \frac{\sum_{G_i} freq-sev \text{ premium} - \sum_{G_i} \text{ premium model} (p = \phi)}{\sum_{G_i} freq-sev \text{ premium}}$$

is plotted and it turns out that the largest part of the observations is situated in the interval [ ]%. Again, it is very analogous for constant $\phi$ or variable $\phi$, apart from the subpopulations around $G_{4000} - G_{4200}$. These consist of BM levels [ ] and [ ] and are consistent with figure 6.8 where we also saw the fluctuations in the difference of constant and variable $\phi$. The (kind of) we saw there around $G_{3900}$ and $G_{4250}$ and around $G_{4100}$ return here: around $G_{3900}$ a [ ] is observed for the tweedievarmodel since at this point, the difference between the tweedievarmodel and the tweediemodel was [ ]. We won’t repeat the analysis for the subpopulations again since they all return (for example $G_{1100}$).
6.6 Tarification in practice

After working with those different premium models, it’s obviously more transparent to use a frequency(-severity) model. Then you know exactly where the numbers come from, how they are estimated; while this is not at all easy for a Tweedie model. As seen from the analysis, the difference between the Tweediemodels and the frequency-severity model is not very large. The computing time for SAS is longer for a Tweediemodel, but in the other case you still have to multiply the frequency and severity estimates, so in total that would not make a difference.

In practice, there are a lot of other factors that have to be considered. It is of course not the goal to blindly apply these models.

- First, there is the problem that some estimates are not accurate enough.
  This can be examined by the coefficient of variation for example; this is defined as $CV = \sqrt{\text{Var}/E}$ or $CV_{\hat{\beta}_i} = SE_{\hat{\beta}_i}/\hat{\beta}_i$. This has to be, in absolute value, as small as possible. In the different models, we found that most values are smaller than .
  For the frequency and severity model we found, as expected, other problem zones. For the frequency the first interaction, BM*age_cont, had a $CV$ for BM levels ; for BM*age this was and for BM*age this was and . For the severity we found that BM levels performed.
for the first interaction, for BM*age it was Also of BM*age had a CV and of age_car*ind_claim_5year. Overall, the severity seems a model than the frequency model, which is naturally the consequence of the smaller dataset to estimate the values from.

For the tweedievar models we found a value ( ) for the estimate of BM*age_cont for BM level . Also for the estimate of BM*age we found values for BM levels (this last only for age equal to ), and for BM*age for age equal to and BM levels . For the tweediemodels we found in general . For BM*age_cont also level gave a value, but for BM*age only level gave a problem for and for BM*age only again for . Finally the level of age_car*ind_claim_5year performed also bad for . The largest value seems to perform than the smaller p value, both for constant and variable φ.

- Second, the tarification factors as resulting from the model may be not continuous or do not evolve smoothly, as a consequence of the lack of accuracy, but also the data that does not behave how you would like it of course. This is a problem in practice since some factors really reflect an increasing (or decreasing) risk. Look at the tarification tables for the BM for example, both for the frequency and freq-sev model, they do not increase necessarily for increasing BM. This is against all common sense of the meaning of the BM and could not be sold.

- Also, the resulting tariff has to be realistic and possible for each risk category. After multiplying some factors for a certain set of characteristics, the premium may be extremely high or low. For the frequency premium for example, one of the premiums (note that there are ) would be circa for a policyholder of years old, with BM , not claimfree in the past 5 years, driving a car of years old, living in zone 7. The same policyholder with the same car, but claimfree and BM , living in zone 1, would pay . So not only are the absolute premiums too cheap or too expensive, but also the relativities, the differences between the possible set of premiums, are too big. This unbalance has to be restored and premiums have to be redivided among the population to create a more realistic, plausible set of premium levels. Insurers would first determine a certain, realistic, premium level, and then lower these premiums by dividing their commercial budget over their portfolio, with certain discounts or rewards.
Finally, the competition is watching. When other premiums in the market tend to be cheaper, you will lose policyholders unless you also lower your tariff. When other premiums are more expensive, you would be a fool not to follow them (a bit) and rise your tarification rates.
Chapter 7

More possibilities and conclusion

In this final chapter, we will first sketch (briefly) possible other models or modelling techniques. Then we will discuss the main results of this thesis, comment shortcomings or possible other methods, and make some conclusions.

7.1 More possibilities

Poisson regression seems the most appropriate distribution to model the claim frequency here, although this seems to be the easiest distribution. Negative binomial distribution is often applied to correct for overdispersion, this distribution is applied in the first section. Other possibilities try to correct for the excess zeros and these are discussed in the second section, namely distributions that arise when trying to capture the very high possibility of observing a zero, like zero-inflated and hurdle models.

7.1.1 Negative binomial distribution

The negative binomial distribution (NB) is also an option; this distribution describes the number of successes in a sequence of independent and identically distributed Bernouilli trials, before a certain number of failures occurs. It is parametrized by $p$, the probability of succes in one trial (so $0 < p < 1$), and $r > 0$, the number of failures. Its mean is then given by $\mu = rp/(1 - p)$ and the variance $Var = rp/(1 - p)^2$, so the extra parameter $r$ reflects a larger variance:

$$Var = \frac{1}{rp} \mu^2 = \mu + \frac{1}{r} \mu^2.$$
Another interpretation is that there is an error term $\epsilon$ included (estimator $\sum_{j=0}^{p} x_j \beta_j + \epsilon$) with $\exp(\epsilon)$ Gamma distributed.

Note that the Poisson distribution can be seen as a limiting case of the NB distribution. When the Poisson parameter (thus the mean or variance of the Poisson distribution) is defined as $\lambda = rp/(1 - p)$, or $p = \lambda/(r + \lambda)$, so that the means of the Poisson and NB distribution are equal, there holds:

$$Pois(\lambda) = \lim_{r \to \infty} NB \left( r, \frac{\lambda}{r + \lambda} \right).$$

With the explanatory variables BM*age_class, BM*age, BM*age, age_car*ind_claim_5year and newzone, we found an AIC score of 5637.21, versus 5638.66 for the Poisson frequency model. The SAS dispersion parameter was estimated at 0.01, this is $1/r$ so that $r$ is estimated as 100. That $r$ would be very small was to be expected as the observed mean frequencies are very small; the NB distribution becomes more asymmetric with small mean and skew to the right for decreasing $r$. Also the parameter estimates are very close to each other. This is mostly explained by the small rescaling of the variance in our dataset, the scale parameter was only around 1.

### 7.1.2 Zero-inflated and hurdle models

Obviously, the possibility of observing zero is very large for the dataset here, recall that a policy with duration one year reported on average 0 claims. This is characteristic for civil responsibility data and is not explained by a certain (combination of) parameter(s). So naturally, models were developed that create a bigger pool of zeros, inherent to the model itself.

**Zero-inflated models**

A first class of models divides the response in two groups $G_1$ and $G_2$ by adding an extra binomial distribution in a first step. The response will be either in $G_1$ with probability $p$ (the zero-inflation probability), where it is automatically zero, or in $G_2$ with probability $q = 1 - p$, where it takes a value from a chosen distribution $X$ (for example Poisson). So the logit model (the logit function is the canonical link function of the binomial distribution) describes the excess zeros.
If $q_0$ denotes the probability of being zero in $G_2$, then the probability of observing a zero response is $p + (1 - p)q_0$ and of observing a non-zero response is $(1 - p)(1 - q_0)$. The mean response is then $(1 - p)\mu < \mu$, with $\mu$ the mean of the chosen distribution $X$.

For example the density for a Poisson distribution with parameter $\lambda$ is given by:

$$
Pr(Y = 0) = (1 - p)\lambda^0/k! \\
Pr(Y = k) = (1 - p)\lambda^k/k! \quad (k > 0)
$$

(7.1)

with the mean and variance then given by

$$
E(Y) = (1 - p)\lambda \\
Var(Y) = E(Y^2) - E^2(Y) \\
= (1 - p)E(X^2) - (1 - p)^2\lambda^2 \\
= (1 - p)\lambda(1 + \lambda) - (1 - p)^2\lambda^2 \\
= (1 - p)\lambda + (1 - p)\lambda^2(1 - (1 - p)) \\
= E(Y) + \frac{p}{1 - p}E^2(Y).
$$

The parameters $p$ and $\lambda$ then depend on two (possibly different) sets of predictors. Among other pioneers of the ZIP distribution, Lambert considered in [7] this model with the logit link function for $p$ and the ln link function for $\lambda$.

In SAS, ZIP models (zero-inflated Poisson) can be easily implemented in the GENMOD statement by choosing `dist = zip’. Other options, in more advanced software versions, are the FMM or COUNTREG statement.

**Hurdle models**

A second class of models involves a **hurdle** that has to be crossed in a first step, to be able to produce a non-zero response. This hurdle is mostly modelled by a binomial distribution: either one crosses the hurdle and a response larger than zero is produced, larger than zero, or one doesn’t cross the hurdle and the response is zero per definition. So the difference with the zero-inflated models is that in this case, only one group produces zeros, the group that doesn’t cross the hurdle. The distribution that produces only non-zeros is called a zero-truncated distribution, for example a zero-truncated Poisson distribution is derived
as:

\[
Pr(Y = 0) = \pi_0 \\
Pr(Y = y) = Pr(Y > 0)Pr(Y = y | Y > 0) \\
= (1 - \pi_0) \frac{Pr(Y = y)}{Pr(Y > 0)} \\
= (1 - \pi_0) \frac{Pr(Y = y)}{1 - Pr(Y = 0)} \\
= (1 - \pi_0) \frac{\lambda^y \exp(-\lambda)}{y!(1 - \exp(-\lambda))}.
\]

Comparing with (7.1), one can see that this is a reparametrization of the zero-inflated Poisson distribution, with \( \pi_0 = p + (1 - p) \exp(-\lambda) \). However different parameters are modelled, \( p \) for the ZIP model and \( \pi_0 \) for the poisson hurdle model, so the models are not at all equivalent [9].

In SAS, hurdle models can be programmed in the statement NLMIXED defining the loglikelihood. Another option is the FMM statement.

**Source of the zeros**

Clearly the zeros in a hurdle model are not random (they don’t cross the hurdle for a certain reason), whereas in the zero-inflated model there are also random zeros in the outcome of the chosen distribution. This subtle difference reflects the source of the zeros: true or structural zeros are actually zeros, meaning that it was possible to have a non-zero, but due to certain circumstances the observation response was zero, excess zeros are zeros that are observed but couldn’t be non-zero. For example [3], if one wants to know on a camping the average number of fish the visitors catch, the true zeros come from people who actually went fishing, the excess zeros come from people who visited the camping but didn’t fish or try to catch one.

The zeros in this dataset, where a policyholder doesn’t report any claim during the insured period, are excess zeros since certain accidents happen through random events or circumstances. For example when a bicycle driver makes an unexpected movement, the driver should be able to stop anytime, but this accident would (mostly) not be directly caused by his driving behaviour or characteristics. It is also possible that there was in fact an accident, but the policyholder didn’t report it.
7.2 Comments and conclusions

Clearly there are many factors that the actuary or product (pricing) manager has to take into account to set up a tarification strategy or model.

- First of all, he needs to know what he’s doing - the theory regarding the possible distributions and used models, interpretation of the output that would be given, modelling techniques of the statistical program, . . . . This takes a lot of time to acquire this knowledge, however there are very interesting books and articles with clear explanations (see the reference list). Moreover the numerous fora on the internet, and help platforms of software are also very helpful.

- Second, he must take decisions about the strategy that will be followed. Certain choices must be made concerning the distribution, significance limits, number of variables, possible interactions. Will the tarification be based on a frequency model, a freq-sev model or a Tweedie model? These are all options that depend on the dataset, the goal and the limitations of the company’s resources. Note that for the segmentation we used a severe significance limit of 5% - this is of course up for discussion.

- Then the dataset needs to be prepared. First a thorough univariate analysis of the possible explanatory variables is necessary, then a bivariate analysis if interactions would be included. Sometimes data cleaning is needed to be able to make accurate conclusions or deductions. The dataset must be understood: why certain trends are visible or not visible, how the variables are related with respect to their frequency but also with respect to the response variables. Segmentation has to be performed if necessary.

- The actuary is then ready for the most exciting part, the modelling. Performance of models needs to be fully understood and tested to make sure that right conclusions can be made upon the performed analysis. The interpretation of the models should be looked into carefully, since each model has its own strengths and weaknesses.

- When the model is developed, this is an excellent guide to build the tariff. The tarification factors should be smoothened, mostly brought closer together so that the portfolio would have a balanced premium distribution among its subpopulations. The final premium may decrease one last time in the commercial process since discounts
are granted to the policyholders directly or to the agents, these may spoil their best
clients with their own budget.

The first part, the theoretical part, is explained in the first two chapters. For completeness,
we also gave a summary (and later on an application) of the Büllmann-Straub estimators
in chapter three. The second part of tarification, the application to the dataset, covers the
other chapters of this thesis. This was done at the best of my abilities, given the short
period and limitations of the SAS version of the company, but it is of course not perfect.
Many things could be done differently, maybe for better of maybe for worse.
As already mentioned, the segmentation is only based on the frequency. This would be
more complete if it was also based on the severity. One could for example perform the
modelling and LSmeans statement in every step both for the frequency and severity model,
put them together and then decide which levels should be grouped. Repeating this step
by step, the segmentation would then result in probably more levels, but no levels would
be joint if it was not conform with both frequency and severity.
Also, the segmentation is only based upon the variable itself, and its relation with the
response variable. No other variables are included - which could certainly be important
when using interactions later on in the premium models. This is tricky since one can not
know beforehand, which interactions will be used, since these tests are performed (mostly)
with segmentation results already. Considering all possible interactions would be a solu-
tion, but not an efficient naturally.
The significance limit for interactions is here 5% while in practice 10 and 20% are widely
used. This would give rise to more possible interactions and hence complicate the choice
of explanatory variables to be included in the model. We could perform the range of
type3tests again for a set of higher significance levels and investigate the influence of this
level on the list of interactions.
Then there is the situation concerning the relation between the BM and ind_claim_5year.
Moreover the BM is not the real BM (this was explained in the first chapter), so possible
correlation with each other or with the response variables is not easy to investigate. It
might be useful to determine a BM in the future with not so many levels, as the 22 levels
now are not handy or useful to work with. Possibly, it could be replaced by a set of indi-
cators, for example the combination of ind_claim_2year, ind_claim_5year, ind_claim_8year
and ind_claim_10year.
The models could be extended with more variables; however to be able to do this, the
portfolio has to be large enough and this would only complicate the tarification. The in-
terpretation covers even more different levels then, and there would be more risk categories
than the 4346 we already have. The analysis could be performed for each year separately however the dataset would have to be large enough. This would not contribute to the time-independency assumption, as one year is still considered with the four seasons and so on, so several years would be better for this assumption. However the policy-independency would be more satisfied then - not completely since a one-year contract is still divided in different parts (so different observation lines) if something changes in cover, premium, co-drivers, and so on. It would certainly be interesting to model for example the evolution in claim frequency over subsequent years. But a tarification model is mostly used for several years, so must be quite robust to yearly changes.

**Different premium models**

In the previous chapter we discussed four general types of premium models: frequency based and frequency-severity based premiums, Tweedie models with constant $\phi$ and variable $\phi$. The first two models are intuitively easier to understand and interpret, since it handles the claim frequency and claim severity. When choosing for Tweedie models, it would be crucial that the actuary first experiences the frequency and severity modelling for developing his insight in the dataset.

With respect to this dataset, it became clear that mostly the BM and the age of the policyholder are important factors in the modeling. As for the modelled premiums, we summarize here the categories with the largest differences. Note that this is often explained by the claim severity that would be relatively lower or higher, generating these differences.

- The policyholders with BM generate a the frequency and freq-sev premiums, and also between the frequency and tweedie premiums. differences are observed for age For the policyholders, the frequency factor is vs for the freq-sev model.

- The same kind of differences is observed for BM regarding the age categories, and for BM and age . Again the frequency estimates are than the freq-sev estimates.

- For BM level and age, these differences are also observed but here the frequency factors are than those of the freq-sev model. This is also the case for BM and age , and BM and age .
7.2 Comments and conclusions

- BM level indicates the differences in tweedievarmodels between the two \( p \)-values for age categories and . Note that for these groups, the estimated frequency factors are the freq-sev factors. Moreover for age between , the frequency factors are than the freq-sev factors, leading to a frequency premium that is than the other modelled premiums.

- BM level produces probably the trends, also the estimates were calculated here. For the age categories ( ) and the (), the frequency factors are than the freq-sev factors - explaining that the frequency premium is than the tweedie(var)premiums. Other age categories display: the frequency factors are than the freq-sev factors. This results in a freq-sev premium or tweedie(var)premium than the frequency premium, and here also the difference in tweediemodels between the two \( p \)-values was .

Thus one should carefully investigate the claim severity, and how this would effect the pure premium estimates. This may be an excellent opportunity to review the current estimation methods of the ultimate cost.

When looking at the difference between the modelled and earned premiums, we find that the following categories are characteristic.

- For age category , policyholders with BM level have modelled premiums that are than the paid premiums.

- This trend continues for BM level for age .

- Also for BM the modelled premiums are than the paid premiums.

It is certainly not a coincidence that these levels, But this is only a part of the explanation of these differences. These correspond in the claim frequency or severity plots, with respect to other BM levels. So in reality these categories, given BM. Given this trend, it may definitely be more useful to create a new ‘BM’ like’ variable with less levels. It would simplify the premium calculation and when done properly, it would give a risk classification that is certainly as good as the old (current) classification.
Future

As for the future, I hope that this company will really explore the possibilities of modelling pure premiums and developing a model-based tarification. A lot of insurers use software provided by specialized (consultant) companies. This software is usually up to date and user-friendly but also very expensive. Sometimes it may be easier to interpret than standard statistical software, because of the lack of knowledge and know-how: the users often don’t know what’s behind the options, output and conclusions. But by research, practice and exercises with statistical software on the data itself, this could be improved in the future. The black-box problem that is mostly attached to this preprogrammed tarification software, would then (certainly partly) be resolved. It would be cheaper also, being independent of this software - maybe not in the beginning when trainings have to be scheduled and paid for, but certainly after a while. Moreover, the general, external pricing software could then be replaced by an internal tarification model, specifically designed for the population, product or strategy of the insurer.

With model-based tarification, it would definitely be more easy to change the tariff regularly, or evaluate it yearly for example. Modelling is a fast and efficient way to indicate trends or evolutions in the (claims) data. Simulations of changes of any kind (tariff, franchise, inflation, exclusion or inclusion of certain subpopulations) can be done rapidly by modelling with statistical programs.

Note that this modelling process can be applied to many types of products or insurance contracts and not only TPL in motor insurance. Poisson modelling could be tested for claims regarding the omnium coverage, or fire insurance contracts. This would be of course an excellent strategy for the insurer: developing a kind of template or general model to estimate pure premium, and calibrate this model to each specific product or population. This makes immediately general and uniform reporting templates or tariff models possible, through programming it in the template model.
Appendices

Appendix A: Additional statistics and variable insights

Zone_cl

Table 7.1: Mean values for zone_cl, the original geographical division.

<table>
<thead>
<tr>
<th>Zone_cl</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>PctSum</td>
<td>dur</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean y</td>
<td>premium</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>costult</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>nclaim</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean n</td>
<td>costult</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Obviously the different zones contain indeed policyholders since the claim frequency with increasing zone, the claim severity. The premium follows the of the pure premium. Notice that zones 3, 4 and 5 sum up to ultimate cost.
7.2 Comments and conclusions

ZoneG

Table 7.2: Explanation of the different levels of zoneG, the percentage of each level in the total duration and the observed mean claim frequency.

<table>
<thead>
<tr>
<th>Level</th>
<th>Explanation</th>
<th>Pct_dur</th>
<th>mean claim frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>unknown</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A01</td>
<td>wealthy professionals</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A02</td>
<td>affluent maturity</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A03</td>
<td>enterprising luxury</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A04</td>
<td>prosperous pensioners</td>
<td></td>
<td></td>
</tr>
<tr>
<td>B05</td>
<td>wealthy families with teenager</td>
<td></td>
<td></td>
</tr>
<tr>
<td>B06</td>
<td>upcoming young families</td>
<td></td>
<td></td>
</tr>
<tr>
<td>B07</td>
<td>rural tranquility</td>
<td></td>
<td></td>
</tr>
<tr>
<td>B08</td>
<td>mature affluent couples</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C09</td>
<td>happily retired</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C10</td>
<td>older blue collar</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C11</td>
<td>young working class</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D12</td>
<td>comfortable skilled workers</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D13</td>
<td>close to retirement</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D14</td>
<td>hardworking entrepreneurs</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D15</td>
<td>commuting white collars</td>
<td></td>
<td></td>
</tr>
<tr>
<td>E16</td>
<td>stable employees in villages</td>
<td></td>
<td></td>
</tr>
<tr>
<td>E17</td>
<td>older rural houses</td>
<td></td>
<td></td>
</tr>
<tr>
<td>E18</td>
<td>renovated family houses</td>
<td></td>
<td></td>
</tr>
<tr>
<td>E19</td>
<td>agricultural community</td>
<td></td>
<td></td>
</tr>
<tr>
<td>E20</td>
<td>isolated rural families</td>
<td></td>
<td></td>
</tr>
<tr>
<td>F21</td>
<td>non-working in social housing</td>
<td></td>
<td></td>
</tr>
<tr>
<td>F22</td>
<td>ageing in working class areas</td>
<td></td>
<td></td>
</tr>
<tr>
<td>F23</td>
<td>workers in older terraces</td>
<td></td>
<td></td>
</tr>
<tr>
<td>F24</td>
<td>italian roots</td>
<td></td>
<td></td>
</tr>
<tr>
<td>G25</td>
<td>modest single living</td>
<td></td>
<td></td>
</tr>
<tr>
<td>G26</td>
<td>employees in terraced housing</td>
<td></td>
<td></td>
</tr>
<tr>
<td>G27</td>
<td>mixed urbanities</td>
<td></td>
<td></td>
</tr>
<tr>
<td>G28</td>
<td>retired flat dwellers</td>
<td></td>
<td></td>
</tr>
<tr>
<td>H29</td>
<td>young urban prosperity</td>
<td></td>
<td></td>
</tr>
<tr>
<td>H30</td>
<td>trendy young singles</td>
<td></td>
<td></td>
</tr>
<tr>
<td>H31</td>
<td>inner city deprivation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>H32</td>
<td>young families in flats</td>
<td></td>
<td></td>
</tr>
<tr>
<td>H33</td>
<td>struggling ethnic families</td>
<td></td>
<td></td>
</tr>
<tr>
<td>U99</td>
<td>unclassified</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### 7.2 Comments and conclusions

#### Code_sport

**Table 7.3:** Mean values for code_sport (95% CI).

<table>
<thead>
<tr>
<th>Code_sport</th>
<th>N</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>PctSum</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean y</td>
<td></td>
<td></td>
</tr>
<tr>
<td>premium</td>
<td></td>
<td></td>
</tr>
<tr>
<td>costult</td>
<td></td>
<td></td>
</tr>
<tr>
<td>nclaim</td>
<td></td>
<td></td>
</tr>
<tr>
<td>upper limit CI</td>
<td></td>
<td></td>
</tr>
<tr>
<td>lower limit CI</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean n costult</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

#### Diesel

**Table 7.4:** Mean values for diesel (95% CI).

<table>
<thead>
<tr>
<th>Diesel</th>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>PctSum</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean y</td>
<td></td>
<td></td>
</tr>
<tr>
<td>premium</td>
<td></td>
<td></td>
</tr>
<tr>
<td>costult</td>
<td></td>
<td></td>
</tr>
<tr>
<td>nclaim</td>
<td></td>
<td></td>
</tr>
<tr>
<td>upper limit CI</td>
<td></td>
<td></td>
</tr>
<tr>
<td>lower limit CI</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean n costult</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

#### Capacity

**Table 7.5:** Mean values for capacity.

<table>
<thead>
<tr>
<th>Capacity</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9(+)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PctSum</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean y</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>premium</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>costult</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>nclaim</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean n</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>costult</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### 7.2 Comments and conclusions

**Ind_omnium**

Table 7.6: Mean values for ind_omnium (95% CI).

<table>
<thead>
<tr>
<th>PctSum</th>
<th>Ind_omnium</th>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>dur</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean y</td>
<td>premium</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>costult</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>nclaim</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>upper limit CI</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>lower limit CI</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean n</td>
<td>costult</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Ind_reduction**

Table 7.7: Mean values for ind_reduction (95% CI).

<table>
<thead>
<tr>
<th>PctSum</th>
<th>Ind_reduction</th>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>dur</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean y</td>
<td>premium</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>costult</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>nclaim</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>upper limit CI</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>lower limit CI</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean n</td>
<td>costult</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Overtake**

Table 7.8: Mean values for overtake (95% CI).

<table>
<thead>
<tr>
<th>PctSum</th>
<th>Overtake</th>
<th>N</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>dur</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean y</td>
<td>premium</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>costult</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>nclaim</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>upper limit CI</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>lower limit CI</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean n</td>
<td>costult</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 7.9: Mean values for private (95% CI).

<table>
<thead>
<tr>
<th></th>
<th>Private</th>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>PctSum</td>
<td>dur</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean y</td>
<td>premium</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>costult</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>nclaim</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>upper limit CI</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>lower limit CI</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean n</td>
<td>costult</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
The distribution of age\_policy for the claimfree policies (figure \ref{fig:age_policy}) has because almost \( \ldots \) of the duration of this subpopulation has age\_policy \( \ldots \). For the other subpopulation, the policies with a claim in the last 5 years, this is \( \ldots \). A share of \( \ldots \) has age\_policy equal to \( \ldots \), whereas the other levels \( \ldots \) each represent about \( \ldots \) of this subpopulation.

**Figure 7.2:** Boxplot of the age of the car (left) and received discount (right) for different levels of newzone (vertical axis).

**Figure 7.3:** Boxplot of the age of the policy (horizontal) for discount (left) and ind\_claim\_5year (right).
Appendix B: SAS code

Univariate statistics and first insights

Summary table with different chosen statistics (here colPctSum and mean) of certain variables (dur, nclaim, ...), given per level of the specified CLASS variable - to obtain the given tables.

PROC TABULATE DATA=DATA.DATATHESIS OUT=WORK.SummaryTables;
VAR nclaim premium premium_real cost costlim costult;
CLASS agecar / ORDER=UNFORMATTED MISSING;
TABLE agecar*
   (ColPctSum Mean)
Mean,
dur
nclaim
premium
cost
costult;
RUN;

Distribution analysis (here of premium) with or without weights.

ODS GRAPHICS ON;

PROC KDE DATA=DATA.DATATHESIS;
UNIVAR premium;
WEIGHT dur;
RUN;

Poisson model with duration as weights, and predicted variable nfreq (= nclaim/dur) to obtain the model for claim frequency and the associated 95% confidence intervals (log as link function and no intercept).

PROC GENMOD DATA=DATA.DATATHESIS PLOTS(ONLY)=NONE;
CLASS agecar;
WEIGHT dur;
MODEL nfreq = agecar / LINK=LOG DIST=POISSON NOINT PSCALE TYPE3;
RUN;
Gamma model with nclaim as weights, and predicted variable severity (= costult/n-claim) to obtain the model for claim severity and the associated 95% confidence intervals (log as link function and no intercept). SAS automatically uses only the observations with costult > 0.

PROC GENMOD DATA=DATA.DATATHESIS PLOTS(ONLY)=NONE;
CLASS agecar;
WEIGHT nclaim;
MODEL severity = agecar / LINK=LOG DIST=GAMMA NOINT PSCALE TYPE3 ;
RUN;

Correlations

Bivariate frequency count (here of age_car and age_policy) with (or without, omitting the WEIGHT statement) weights is possible with the KDE statement.

PROC KDE DATA=DATA.DATATHESIS;
BIVAR agecar age_policy;
WEIGHT dur;
RUN;

However we made figures with the SGPLOT statement because the variation is more clear given the boxplot.

PROC MEANS DATA=DATA.DATATHESIS MISSING NWAY NOPRINT;
CLASS DISCOUNT NPOL;
VAR NCLAIM DUR;
OUTPUT OUT=DUM SUM= / AUTONAME;
RUN;

PROC SGPLOT DATA=DUM;
HBOX DISCOUNT / CATEGORY=NPOL FREQ=DUR_SUM ;
RUN;
7.2 Comments and conclusions

Segmentation of variables

Construction of Poisson model with categorical variable (here for example age) to obtain the parameter estimates; then the LSmeans are computed. Note that the reference level in the Poisson model will be determined as the level with the most observations by the statement `DESC ORDER=freq`.

```
PROC GENMOD DATA=DATA.DATATHESIS;
CLASS age (DESC ORDER=freq);
WEIGHT dur;
MODEL nfreq=age / D=POISSON PSCALE TYPE3;
ODS OUTPUT PARAMETERESTIMATES=PARMS;
ODS OUTPUT MODELFIT=MODELFIT;
STORE AUTO_AGEW;
RUN;

PROC PLM SOURCE=AUTO_AGEW;
SHOW FIT;
LSMEANS age /DIFF;
ODS OUTPUT DIFFS=DIFFAGEW;
RUN;

DATA DATA.DATATHESIS;
SET DATA.DATATHESIS;
agew4class = 0;
if agew4 = then agew4class = 1;
if agew4 = then agew4class = 2;
if agew4 = then agew4class = 3;
if agew4 = then agew4class = 4;
if agew4 = then agew4class = 5;
if agew4 = then agew4class = 6;
if agew4 = then agew4class = 7;
if agew4 = then agew4class = 8;
if agew4 = then agew4class = 9;
age = 0;
if then age = 1;
age = 0;
```
if then age = 1;
if then age = 1;
run;

PROC GENMOD DATA=DATA.DATATHESIS;
CLASS age (DESC ORDER=freq);
CLASS age (DESC ORDER=freq);
WEIGHT dur;
MODEL nfreq=agew4class age age / D=POISSON PSCALE TYPE3;
RUN;

Interaction terms
The command type3 gives the results of type3 significance tests.

PROC GENMOD DATA=DATA.DATATHESIS;
CLASS variable1 (DESC ORDER=freq);
CLASS variable2 (DESC ORDER=freq);
WEIGHT dur;
MODEL nfreq = variable1*variable2 variable1 variable2 / D=POISSON PSCALE TYPE3 ;
RUN;

Model building
The subsequent frequency models for stepwise forward selection are already included in chapter six.
Tweedie modelling using first the model with constant $\phi$, then the model with variable $\phi$, which is defined in a DATA statement before the GENMOD statement to spread the calculations over two statements. As also stated in the text, the dataset ‘Modelpp’ contains already the variables

- ‘freqpred’: estimated frequency from the frequency model with the variables as below,
- ‘sevpred’: the estimated severity from the severity model with those variables,
- ‘pure_premium’: costult/dur.
The OUTPUT statement then simply adds the new predictions to this dataset.

PROC GENMOD DATA=DATA.MODELPP;
CLASS agecar4 (DESC ORDER=freq);
CLASS agew4 (DESC ORDER=freq) age (DESC ORDER=freq);
CLASS age (DESC ORDER=freq) ind_claim_5year (DESC ORDER=freq);
p = ;
phi = ;
p1 = 1-p;
p2 = 2-p;
mu = _MEAN_
;y = _RESP_
;w=dur;
VARIANCE var=(phi*(mu**p))/w;
if( y=0 ) then d = 2*w*(mu**p2)/p2;
else if( y>0 ) then d = 2*w*(y**p2 - p2*y*mu**p1 + p1*mu**p2)/(p1*p2);
DEVIANCE dev=d;
MODEL pure_premium=BM*age\_cont BM*age BM*age\_car*ind\_claim\_5year newzone / LINK=LOG TYPE3 MAXITER=150;
ods output ParameterEstimates=parms;
ods output Modelfit=Modelfit;
OUTPUT out=data.modelpp pred=tweediepredpp;
RUN;

DATA DATA.MODELPP;
SET DATA.MODELPP;
phi = (sevpred**())/((freqpred**())*);
RUN;

PROC GENMOD DATA=DATA.MODELPP;
CLASS agecar4 (DESC ORDER=freq);
CLASS agew4 (DESC ORDER=freq) age (DESC ORDER=freq);
CLASS age (DESC ORDER=freq) ind_claim_5year (DESC ORDER=freq);
p = ;
p1 = 1-p;
p2 = 2-p;
mu = _MEAN_; 
y = _RESP_; 
w=dur; 
phi= phi;
VARIANCE var=(phi*(mu**p))/w;
if( y=0 ) then d = 2*w*(mu**p2)/p2;
else if( y>0 ) then d = 2*w*(y**p2 - p2*y*mu**p1 + p1*mu**p2)/(p1*p2);
DEVIANCE dev=d;
MODEL pure_premium=BM*age\_cont BM*age\_car BM*age\_newzone / LINK=LOG TYPE3 MAXITER=150;
ods output ParameterEstimates=parms;
ods output Modelfit=Modelfit;
OUTPUT out=data.modelpp pred=predpp;
RUN;

Appendix C: SAS output

The following figures are the exact output as given by SAS gives when running the statement:

PROC GENMOD DATA=DATA.DATATHESIS;
CLASS age\_car bm4 (DESC ORDER=freq); 
CLASS age\_w4 (DESC ORDER=freq) age\_newzone (DESC ORDER=freq); 
CLASS age\_w (DESC ORDER=freq) ind\_claim\_5year (DESC ORDER=freq); 
CLASS zoneg (DESC ORDER=freq) pow6 (DESC ORDER=freq); 
CLASS diesel (DESC ORDER=freq) npol (DESC ORDER=freq); 
WEIGHT dur;
MODEL nfreq=agew4class*bm4 age\_w4*bm4 age\_newzone*bm4 age\_car4*ind\_claim\_5year newzone / D=POISSON PSSCALE TYPE3 ; 
RUN;
7.2 Comments and conclusions

Figure 7.4: Output part 1: model information and class level information.
Figure 7.5: Output part 2: goodness of fit criteria and parameter estimates analysis.
7.2 Comments and conclusions

Figure 7.6: Output part 2 (2): parameter estimates analysis.

Figure 7.7: Output part 2 (3): parameter estimates analysis.
Figure 7.8: Output part 2 (4): parameter estimates analysis.

Figure 7.9: Output part 3: type3 analysis.
Index

da posteriori variables, 3
a priori variables, 2
age (segmentation), 72
age (variable), 37, 45, 63
age_car (segmentation), 78
age_car (variable), 37, 43, 62
age_policy (variable), 37, 44
AIC, 30
AICC, 30
BIC, 30
BM (segmentation), 75
BM (variable), 39, 58, 67
bonus malus, 4
BS estimator, 33, 88
BS estimator (zoneG), 85
budget (variable), 37, 56, 61, 66
CAIC, 30
claim frequency, 6
claim severity, 6
compound Poisson distribution, 12, 27
confidence interval, 18
cost (variable), 38
costlim (variable), 38
costult (variable), 38
cover (variable), 37, 55
credibility estimator, 32
cumulant function, 15
deviance, 30
diesel (variable), 38, 53
dispersion parameter, 15, 19
dur (variable), 36
estimators, estimated regression coefficients, 17
excess zeros, 10, 125
experience rating, 3
exponential dispersion family, 15
exposure, 6
Fisher information matrix, 18
frequency model, 39, 89
frequency premium, 101
frequency-severity premium, 105
gamma distribution, 12
gamma model, 22
GLM, 15
goodness of fit measure, 30
homogeneity assumption, 7
hurdle model, 124
ind_claim_5year (variable), 38, 57, 67
independency assumption, 7
interaction, 4, 80
link function, 16
loglikelihood, 15
INDEX

LS means, 71
mean, weighted, 7
mixed Poisson distribution, 10
model selection, forward/backward, 88
multiplicative tariff, 21, 22

nclaim (variable), 38
negative binomial distribution, 122
newzone (linearity), 79
newzone (variable), 37, 48, 64
npol (variable), 38, 58

offset, 23

Pearson's chi-square statistic, 20
Poisson distribution, 9
Poisson model, 20
pow (segmentation), 76
pow (variable), 37, 53, 61, 68
premium (variable), 36
pure premium, 6, 101

reference level, 4
regression model, 14
reproductive distribution, 21
risk classification, 2

score function, 16
segmentation, 71
severity model, 39, 98

TPL, 2
Tweedie distribution, 26
Tweedie model, 26, 111

variance, weighted, 7

weight, 24

zero-inflated model, 123
ZIP model, 123
zoneG (variable), 37, 50, 65
Bibliography


List of Figures

4.1 The difference of a non-weighted (left figure) versus a weighted (right figure) analysis for premium ............... 42
4.2 Frequency histogram of the ultimate cost. ...................... 42
4.3 Average claim frequency and 95% CI (dotted lines) for the age of the vehicle. .................. 43
4.4 Average claim severity and 95% CI (dotted lines) for the age of the vehicle. .................. 43
4.5 Average pure premium and 95% CI (dotted lines) for the age of the vehicle. .................. 44
4.6 Weighted frequency histogram of age of the policyholder. ...................... 45
4.7 Average claim frequency and CI for age. ...................... 46
4.8 Average claim severity and CI for age. ...................... 46
4.9 Average pure premium, and CI, and premium for age. .................. 47
4.10 Average claim frequency and 95% CI for newzone. A linear trend is added to fit the average claim frequency, this line has equation claim frequency = \( \text{newzone} \) (computed with Excel). .................. 48
4.11 Average claim severity and 95% CI for newzone. .................. 49
4.12 Average pure premium, and CI, and retained premium for newzone; linear trend pure premium = \( \text{newzone} \) (the observed premium for \( U99 \) is \( \text{newzone} \)). .................. 49
4.13 Share in total sum of premium (left) and ultimate costs (right) for different levels of newzone. .................. 50
4.14 Average claim frequency and 95% CI with respect to zoneG. .................. 51
4.15 Average claim severity and 95% CI with respect to zoneG (no costs were observed for \( U99 \) so this zone is not displayed). .................. 51
4.16 Average pure premium, and CI, and retained premium with respect to zoneG (the observed premium for \( U99 \) is \( \text{zoneG} \)). .................. 52
4.17 Share of different levels of zoneG in total sum of premium (left) and ultimate costs (right). .................. 52
4.18 Average claim frequency (and CI) for the power of the vehicle, limited to the range of 30 to 150. .................. 54
4.19 Average claim severity (and CI) for the power of the vehicle, limited to the range of 30 to 150. 54
4.20 Average pure premium (and CI) and retained premium for the power of the vehicle, limited to the range of 30 to 150. 55
4.21 Mean claim frequency, severity, premium and pure premium for various levels of discount (intervals of budget). 56
4.22 Share in total sum of premium and ultimate cost for different levels of npol. 58
4.23 Mean claim frequency (and CI) for BM. A linear curve to fit the predictions (computed with Excel) is added with equation claim frequency = BM * BM + . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 59
4.24 Mean claim severity (and CI) for BM; linear curve with equation claim severity = BM * BM + . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 60
4.25 Mean pure premium (and CI) and retained premium for BM. 60
4.26 Boxplot of age_car (horizontal) for different levels of discount (left) and power (right). Note that a segmentation is used here for power that will be explained in the next chapter; to deliver a distribution analysis of age_car among the power, the original levels of power would be far too many. 62
4.27 Boxplot of age_car (horizontal) for different categories of cover (left) and different levels of age (right). 63
4.28 Boxplot of the age of the policyholder (horizontal) for ind_claim_5year (left) and boxplot of age (horizontal) for different levels of BM (right). 64
4.29 Boxplot of age_policy (left) and age (right) for the different levels of newzone (vertical axis). 65
4.30 Boxplot of ind_claim_5year (left) and newzone (right) for different categories of zoneG (vertical axis). 66
4.31 Boxplot of discount (horizontal) for different categories of cover (left) and different levels of overtake (right). 67
4.32 Boxplot of npol (horizontal) for different categories of ind_claim_5year (left) and BM (right). 68
4.33 Boxplot of the power of the vehicle (horizontal) for different categories of code_sport (left) and private (right). 69

5.1 Illustration of the small difference between the model from the fourth step (Est, L, U) with 9 levels (estimates) and the corrected model with 4 estimates - continuous with two dummy variables. 74
5.2 Model0 with range from 17 to 85 and its Est, L, U together with the final model - continuous with two dummy variables. ............................................. 74
5.3 Model0 with 23 levels and its Est, L, U together with the third segmentation model .......................................................... 76
5.4 The observed claim frequency for the 61 power levels from model0 with CI limits and estimates for segmentation step 4 with 6 levels. .......... 77
5.5 The observed claim frequency for age_car, the 19 levels from model0 with CI limits and estimates for segmentation 3 with 5 levels. .............. 78
5.6 Different models containing newzone: Est, L and U are obtained from the categorical model with 9 (original) levels. The continuous model (orange line) is equally good as the continuous model with two corrections (black dotted line). ................................................................. 79
5.7 Illustration of interaction effect (blue and red line) and global (only) effect of newzone and ind_claim_5year (dotted lines). ...................... 85
5.8 Illustration of credibility theory for zoneG (the credibility factor is the only variable that refers to the right vertical axis). ....................... 86
6.1 The relative difference of the modelled frequency premium with the earned premium. From left to right: age, age, age_class. ..................... 104
6.2 The relative difference of the modelled frequency premium with the earned premium. Left BM and right age_car. ............................. 105
6.3 The relative difference of the modelled freq-sev premium with the earned premium. Left: ind_claim_5year, right: BM. ......................... 107
6.4 The relative difference of the modelled frequency premium with the freq-sev premium. ............................................................. 108
6.5 The relative difference of the modelled premiums with the earned premiums (blue for the frequency premium and green for the frequency-severity premium, values referring to the left vertical axis). These values are \(d_i\), defined above and can be read on the left y-axis with range \([-\%]%\). The share in total duration is also given (red, values referring to the right vertical axis). ......................................................... 110
6.6 The relative difference between the two \(p\) values of the modelled premiums of the tweediemodels (with constant \(\phi\), in blue) and of the tweedievarmodels (variable \(\phi\), in red). ................................................................. 114
6.7 The relative difference of the modelled premiums with constant \(\phi\) and variable \(\phi\) for \(p\) equal to  (blue) and  (red). ....................... 115
LIST OF FIGURES

6.8 The relative difference of the earned premiums and the modelled premiums for $p$ equal to: the tweediemodel in red and tweedievarmodel in blue. 116

6.9 The relative difference of the frequency premiums and the tweedie(var)premiums for $p$ equal to: the tweediemodel in red (almost not visible) and tweedievarmodel in red. 117

6.10 The relative difference of the freq-sev premiums and the tweedie(var)premiums for $p$ equal to: the tweediemodel in blue and tweedievarmodel in red. 119

7.1 Boxplot of the age of the policyholder (horizontal) for different categories of age_policy (left) and different levels of npol (right). 135

7.2 Boxplot of the age of the car (left) and received discount (right) for different levels of newzone (vertical axis). 136

7.3 Boxplot of the age of the policy (horizontal) for discount (left) and ind_claim_5year (right). 136

7.4 Output part 1: model information and class level information. 143

7.5 Output part 2: goodness of fit criteria and parameter estimates analysis. 144

7.6 Output part 2 (2): parameter estimates analysis. 145

7.7 Output part 2 (3): parameter estimates analysis. 145

7.8 Output part 2 (4): parameter estimates analysis. 146

7.9 Output part 3: type3 analysis. 146
List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>Important ratios</td>
<td>7</td>
</tr>
<tr>
<td>1.2</td>
<td>Key variables statistics over the whole dataset, and the part with non-zero costult</td>
<td>41</td>
</tr>
<tr>
<td>1.3</td>
<td>Frequency of the levels of age_car with respect to the duration</td>
<td>44</td>
</tr>
<tr>
<td>1.4</td>
<td>Age of the policy, limited to 5</td>
<td>45</td>
</tr>
<tr>
<td>1.5</td>
<td>Newzone, the extended geographical division</td>
<td>48</td>
</tr>
<tr>
<td>1.6</td>
<td>Different coverage for the policy</td>
<td>55</td>
</tr>
<tr>
<td>1.7</td>
<td>Claimfree indicator of the five years preceding the policy</td>
<td>57</td>
</tr>
<tr>
<td>1.8</td>
<td>Number of policies by the policyholder</td>
<td>58</td>
</tr>
<tr>
<td>2.1</td>
<td>Variables which will be possibly used and won’t be used for certain in the premium modelling process</td>
<td>71</td>
</tr>
<tr>
<td>2.2</td>
<td>Consecutive steps (or segmentations) for age (Levels denotes the number of levels). The last line corresponds to the final, chosen segmentation as explained in the text</td>
<td>73</td>
</tr>
<tr>
<td>2.3</td>
<td>Consecutive steps (or segmentations) for BM</td>
<td>75</td>
</tr>
<tr>
<td>2.4</td>
<td>Consecutive steps (or segmentations) for pow</td>
<td>77</td>
</tr>
<tr>
<td>2.5</td>
<td>Consecutive steps (or segmentations) for age_car</td>
<td>78</td>
</tr>
<tr>
<td>2.6</td>
<td>Linearity of newzone</td>
<td>79</td>
</tr>
<tr>
<td>2.7</td>
<td>Possible interaction types: two categorical variables (levels denoted by capital letters) with 9 estimates, a continuous and a categorical variable with 6 estimates, and two continuous (levels denoted by numbers) with 3 estimates (α for the interaction, β and γ for the variables alone). Note that there is always a parameter that can be left out because a certain level is chosen as reference level</td>
<td>81</td>
</tr>
<tr>
<td>2.8</td>
<td>Significant (on a 5% level) interaction terms in the frequency and severity model</td>
<td>83</td>
</tr>
</tbody>
</table>
### LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.9</td>
<td>Interaction terms where the algorithm did not converge because of too many levels containing few or none observations (only occurred for the frequency model).</td>
</tr>
<tr>
<td>6.1</td>
<td>Possible terms in the frequency and severity model.</td>
</tr>
<tr>
<td>6.2</td>
<td>Single factor frequency models, their AIC value and their significance (or p-value), ordered with respect to AIC.</td>
</tr>
<tr>
<td>6.3</td>
<td>Subsequent frequency models for stepwise forward selection, their AIC value and their significance (or p-value). The SAS code is included in the next subsection.</td>
</tr>
<tr>
<td>6.4</td>
<td>Subsequent frequency models using stepwise selection when allowing variables to occur multiple times, their AIC value and their significance (or p-value).</td>
</tr>
<tr>
<td>6.5</td>
<td>Subsequent frequency models for forward selection (using each variable only once), their AIC value and their significance (or p-value).</td>
</tr>
<tr>
<td>6.6</td>
<td>Single factor severity models, their AIC value and their significance (or p-value), ordered with respect to AIC.</td>
</tr>
<tr>
<td>6.7</td>
<td>Subsequent severity models for stepwise forward selection (using each variable only once), their AIC value and their significance (or p-value).</td>
</tr>
<tr>
<td>6.8</td>
<td>Subsequent severity models for stepwise forward selection (using variables multiple times), their AIC value and their significance (or p-value).</td>
</tr>
<tr>
<td>6.9</td>
<td>Factors in the (frequency-based) tarification for the interaction age\textunderscore class\textasteriskcentered BM, age\textasteriskcentered BM and age\textasteriskcentered BM. Underlined factors are either or .</td>
</tr>
<tr>
<td>6.10</td>
<td>Factors in the (frequency-based) tarification for newzone (continuous, so the factors are powers of the estimate ).</td>
</tr>
<tr>
<td>6.11</td>
<td>Factor in the (frequency-based) tarification for newzone (continuous, so the factors are powers of the estimate ).</td>
</tr>
<tr>
<td>6.12</td>
<td>Factors in the freq-sev tarification for the interaction age\textunderscore class\textasteriskcentered BM, age\textasteriskcentered BM and age\textasteriskcentered BM.</td>
</tr>
<tr>
<td>6.13</td>
<td>Factors in the freq-sev tarification for the interaction age\textunderscore car\textasteriskcentered ind\textunderscore claim\textasteriskcentered 5year.</td>
</tr>
<tr>
<td>6.14</td>
<td>Factor in the freq-sev tarification for newzone (continuous, so the factors are powers of the estimate ).</td>
</tr>
<tr>
<td>6.15</td>
<td>Goodness of fit statistics for the four different Tweedie models.</td>
</tr>
<tr>
<td>7.1</td>
<td>Mean values for zone\textunderscore cl, the original geographical division.</td>
</tr>
<tr>
<td>Table</td>
<td>Description</td>
</tr>
<tr>
<td>--------</td>
<td>-----------------------------------------------------------------------------</td>
</tr>
<tr>
<td>7.2</td>
<td>Explanation of the different levels of zoneG, the percentage of each level in the total duration and the observed mean claim frequency.</td>
</tr>
<tr>
<td>7.3</td>
<td>Mean values for code_sport (95% CI).</td>
</tr>
<tr>
<td>7.4</td>
<td>Mean values for diesel (95% CI).</td>
</tr>
<tr>
<td>7.5</td>
<td>Mean values for capacity.</td>
</tr>
<tr>
<td>7.6</td>
<td>Mean values for ind_omnium (95% CI).</td>
</tr>
<tr>
<td>7.7</td>
<td>Mean values for ind_reduction (95% CI).</td>
</tr>
<tr>
<td>7.8</td>
<td>Mean values for overtake (95% CI).</td>
</tr>
<tr>
<td>7.9</td>
<td>Mean values for private (95% CI).</td>
</tr>
</tbody>
</table>