Travel Insurance Claims modeling using Statistical Learning

Mitacs Acceleration Research Project

Caroline Leboeuf
Départment d'Informatique et de Recherche Opérationnelle
Université de Montréal

Presented to

Professor Philippe Gagnon
Département de Mathématiques et de Statistique
Université de Montréal

Mr Michel Hébert
Vice-President, Research and Development
Optimum Réassurance Inc.

September 2020

Abstract

Keywords: *data science, generalized linear models, machine learning, model selection, predictive analytics, statistical learning, travel insurance, variable selection*

Contents

1 Introduction

Insurance risk assessment has long been modelled by actuaries using techniques combining concepts from actuarial science, probability, statistics, finance and economy. During the last decades, important improvements in computer performance, advances in the field of machine learning and a boost of access to data-driven information has led to an increased interest in adopting alternative risk modeling. Predictive analytic, defined as a set of statistical learning techniques for predicting trends and outcomes, is now known as a promising path for insurance R&D, including for marketing, underwriting, fraud detection, pricing and reserve valuation.

Property & Casualty (P&C) insurers were leaders in the early use of predictive analytic models given the large amount of claims for some lines of business like auto insurance, and additional insights gained from new technologies like the Internet of Things (IoT). The Canadian Life & Health insurance industry has nevertheless begun introducing such techniques to its business operations, and a general enthusiasm towards R&D projects has helped access to literature for applying theoretical principles specifically to life and health insurance data.

Travel insurance is a distinct line of business which can be found both as an individual and as a group health insurance coverage. Claims are characterized by low incidence frequency, but a heavily skewed severity tail for high claims given the potential for extremely high medical expenses for out of country emergencies. Premiums are therefore usually small, but high volatility of claims for more severe incidents requires precautions for modeling future claims. The lack of literature regarding theoretical applications to travel claims combined with the simplicity of the data available specific to this line of business leads to great opportunities for developing statistical models. Furthermore, increased democratization of travelling is a source of motivation for refining methods of risk assessment.

This report is intended to make use of travel claims data to apply statistical learning methods like Generalized Linear Models (GLMs) to build a predictive model. It also serves as a reference to improve understanding of this small line of business and enable improved rating and pricing calculations. More broadly, the resulting models and the associated methods are a good starting point for extrapolating to more complex insurance products and can be useful for other Life & Health insurance products. All data analysis and visualizations were completed using both Microsoft Excel and the software R.

2 Data

The predictive model presented in this report is based on a regression using Generalized Linear Models (GLM). A good regression model requires a clean and uniform database to improve the significance of explanatory variables and the fitting performance. In this section, we present the details relating to the severity dataset that was used for the regression model and that serves as the foundation for this analysis. First, we describe in Section 2.1 the travel insurance market and how it defines the data and the available information used for modeling. In Section 2.2, we present the various data sources and how they are structured. Given the variability of data sources, a significant step in the research project is to pre-process and clean the dataset, which is explained in Section 2.3. Finally, we present a summary of the available variables and preliminary data analysis in Section 2.4.

2.1 Travel Insurance

Canadian insurance companies started offering travel coverage in the early 90's, while government reduced the medical coverage for out-of-country emergencies. Today, the government covers approximately 7% of travel medical claims. Given this minimal coverage, additional coverage is highly recommended when planning a trip abroad (Frank, 2016).

Travel insurance coverage can be found as a standalone individual insurance product. It can also be part of a group insurance provided by an employer. The coverage includes medical claims and sometimes baggage loss or trip cancellation fees. This study focuses specifically on medical claims.

In group insurance, employees are entitled to a group insurance coverage, and the premium payment is usually split with their employer. Travel insurance can be part of the coverage which most commonly includes disability, life and other health insurances. Each insured employee is given a single certificate number, to which one or multiple additional insureds can be added. An additional insured is known as a dependant and could be a spouse or a children. Therefore, different types of protections are available, such as single, family, spouse or single parent. Insurance policy systems usually aggregate information like the date of birth, the employment status, the salary, the occupation, the maximum amount and duration of the coverage and the postal code, for both the main insured and the dependants. In this case, the list of insureds includes the list of all employees at risk of making a claim, whether or not a trip is planned, and no information regarding the trip is available prior to a claim.

Another type of travel insurance available on the market is the individual travel insurance. It acts differently because a purchase is directly linked to a trip. For this product, all information about a planned trip like the destination, the trip duration or the premium are available when the insurance policy is issued. Furthermore, medical underwriting can be performed, which adds key information for the risk assessment of an insured. The insurance is valid for the specified period of the coverage, which is at most the duration of the trip. Some products are on an annual basis, and others are purchased on a per-trip basis. Given the simplicity of the group insurance coverage, we focus this research report on the analysis of the severity of group insurance. That being said, results shown in this report underline the limitations caused by a lack of information about a trip linked to a claim, and will highly affect the performance of the predictive model.

2.2 Data Collecting

Optimum Réassurance Inc. is a reinsurance company which is regarded as a leader in the Canadian travel insurance industry with several long-time specialists, holding a significant part of the reinsurance market. Given their privileged position, the R&D team suggested to their insurance partners to contribute to a common database. The aim of this project is to better assess the underlying risk of this line of business, and potentially improve rating techniques while gaining knowledge of the most relevant predictive analytic and statistical learning techniques. There are three participating partners, and they will be referred to as Client 1, Client 2 and Client 3.

Predictive analytic techniques are used to model the incidence and severity of an insurance product. The incidence model seeks to predict the risk that an insured individual with some given characteristics will make a claim or not, without regard to the amount of the claim. The severity model will, instead, focus on the prediction of the amount of a claim. The two models are very distinct, with the latter seeking a probability $p \in [0, 1]$, and the former predicting an amount $y \in \mathbb{R}^+$. They each require different model parametrizations, and the target response variable of interest is linked to distinct variables. For example, the proximity of an insured to an airport might affect its propensity to travel and therefore increase the risk of making a claim, but it might not necessarily affect the amount of the claim. This report will focus strictly on a *Severity* model and will act as the starting point for an incidence and severity model in group travel insurance. Results will be useful to compare how it differs from the severity of individual travel insurance products.

The participating insurers provided two types of datasets, which is the foundation of the predictive model. The first dataset consists of their list of insured employees with the details of their coverage. These are the available information for actuarial pricing and can be used in a predictive model. The second is a list of all bills with paid medical claims and the details of the related insured event. These information are only available once a claim is made, and cannot be used in a predictive model, except for the amount of the claim which is used as the response variable in a severity model.

All three participating insurers provided the full list of the main insureds (the employees) with the addition of the list of dependants with key information such as date of birth, gender, etc... Unfortunately, including the dependants details in the predictive model required complex data manipulations, and it was decided to only include details linked to the main insured for each certificate.

Table 1 and Table 2 display the main characteristics of the received datasets. The available fields will be used for the explanatory variables in the regression model.

Table 1: List of Insureds file

Partner	Number of Insureds	Number of Fields	Period
Client 1	661,637	39	2016-12 to 2018-12
Client 2	852,592	32	2014-12 to 2018-12
Client 3	2,046,460	13	2014-12 to 2018-12

Table 2: List of Billings file

Partner	Number of Bills	Number of Fields	Period
Client 1	28,683	47	2016-01 to 2019-10
Client 2	34,312	30	2014-01 to 2018-12
Client 3	127,022	20	2015-01 to 2019-12

2.3 Data Processing

A severity dataset was built using both datasets described in the previous section. In the predictive model, the target response variable is the total claim amount paid by the insurance company for each event covered under the group policy. For this reason, all medical bills related to a single event had to be summed up together to obtain a total claim amount. For example, if an insured individual has a car accident during a trip, then all bills including fees for a stay at the hospital, drug prescriptions and ambulance transportations must be aggregated to result in the total claim amount. The severity dataset therefore consists of the list of all paid claims with the related information of the certificate which made a claim (date of birth, gender, occupation, status, etc...) and the details of the event (medical diagnosis, trip destination, date of the event, period of medical services, etc...). The details of the certificate were taken from the *List of Insureds* file, while the details of the event and the bills amount were taken from the *List of Billings* file. Given that the certificate number for each claim was available, the correspondence between the two datasets was fairly simple. That being said, approximately 10% of the claims did not have a matching certificate number in the *List of Insured* file. The reason for this is that participating insurers provided a *photo* of their list of insureds at the end of the year for a period of 3-4 years, and does not include employees that both started and ended their insurance coverage during the year. Therefore, these claims were removed from the severity dataset, and an incidence analysis should account for this reduction of claims.

Additional fields were added to the dataset for more complex analysis. This includes the type of region linked to the 3 first digits of the postal code (rural or urban), the distance to the nearest airport and to the nearest U.S. border, the province of residence, the Standard Industrial Classification (SIC) linked to the type of occupation, the maximum duration of the trip and the maximum amount of coverage.

Two main challenges appeared in the data processing step. First, claims from the *List of Billings* file were split on a billing basis. A unique identification number was created to group all bills from a specific event using the combination of the certificate number, date of birth, medical diagnosis and medical service date. Client 1 is the only partner who provided an event identification number, which made the aggregation task very simple. The second challenge was to match the correct row from the *List of Insureds* file to each claim. The certificate number served as the main link, but one certificate number can appear in multiple file dates. Therefore, all claims with a date of event between June of year X and June of year X+1 were linked to the corresponding certificate from the file of end of year X. For example, all claims which occurred from June 2017 to May 2018 were linked to the corresponding certificate in the *List of Insureds* file of December 2017. In this case, if a claim is made on August 2019 and the list of insureds stops at December 2018, then this claim will be excluded from the dataset.

Data cleaning consisted of several data manipulations meant to standardize the fields values. For example,

unavailable information were converted to "NA", and typos were removed to ensure proper classification for categorical fields. This step required a substantial amount of work to improve the uniformity of the fields among the three clients and to detect data anomalies.

The final severity dataset has a total of $25,269$ rows, and includes all claims from the three participating partners which have a match in the *List of Insureds* file. Table 3 displays the available fields from the dataset, while table 4 lists the specific fields which will be used as explanatory variables in the predictive model described in the next sections. Unfortunately, not all of the original fields can be used in a predictive model, since some of them are only known once a claim is made which is why there are very few explanatory variables. Nevertheless, the original fields were useful to detect covariance between variables and understand how a model fit changes when we include them. For example, including a binary destination factor *United States versus Other* severely impacted the value of the regression coefficient linked to the province of residence. Such results underline the covariance between the province and the destination, a key consideration for actuaries.

Table 3: Fields description from the severity dataset

Category	Field	Values
Identification	Certificate Number	
	Group Number	
	Employer Number	
Insured	Gender	Female or Male
	Date of Birth	
	Age	16 to 100
	Status	Active or Retired
	Salary	10,000$ and more
	Occupation	Administration, Education...
Coverage	Type	Single, family, couple or single parent
	Maximum amount	1M, 5M...
	Maximum duration	30 days, 182 days...
Geography	Postal Code (FSA)	H2T XXX, A1Y XXX...
	Region Type	Urban or Rural
	Province	Quebec, Ontario, Alberta...
Trip	Start and End of Trip	
	Destination	United States, France...
Claim	Event Number	
	Benefit code	Hospital, Drug Prescription...
	Medical Diagnosis	Flue, Infection...
	Date of claim	
	Start and End of Medical Services	
	Claim Amount	1$ and more

Table 4: Explanatory Variables

Field	Values	Type
Age	Age of insured	Real-valued
Company	Client 1, Client 2 or Client 3	Categorical
Gender	Female or Male	Binary
Protection Type	Couple, Family, Single or Single Parent	Categorical
Province	British-Columbia, Ontario, Quebec, etc...	Categorical
Salary	10,000$ and more	Real-valued
Status	Active or Retired	Binary
Year	Year of claim event	Real-valued

2.4 Data Description

2.4.1 In Force

A policyholder who has paid its premium and is eligible to the coverage underlined in the policy has an *In Force* insurance. For group travel insurance, all employees who adhere to the coverage made available from their employer are eligible and constitute the list of insured individuals at risk, and will be referred to as the *In Force*. Table 5 to Table 9 summarize how the group of *In Force* is distributed among different variables of interest.

Table 5: Company breakdown

Client 1	Client 2	Client 3
17%	24%	60%

Table 6: Period breakdown

Compagnie	2014-12-31	2015-12-31	2016-12-31	2017-12-31	2018-12-31
Client 1	-	-	31%	33%	36%
Client 2	25%	25%	24%	26%	-
Client 3	-	22%	25%	25%	28%

2.4.2 Severity

The variable of interest in the severity analysis is the total paid claim amount by the insurer associated to a single event. This amount includes the accepted claim amount to be paid, with the addition of fees and taxes. Refunds are currently excluded from the analysis.

Table 10 presents statistics describing the claim amounts. The median of 301.40$ is significantly lower than the mean claim amount of 2,373.70$. Furthermore, the 3rd quantile is of 955.20$ while the maximum claim amount is over 1M.

One of the main issues underlined during claims analysis and modeling is the extreme spread of claims. In this report, we are mainly interested in the more simple GLM models, but a mixture model appeared as

Table 7: Gender (left) and Region Type (right) breakdown

Female	Male
46%	54%

Rural	Urban
14%	86%

Table 8: Type of Coverage (left) and Insured Status (right) breakdown

Single	Couple	Family	Single Parent
46%	16%	36%	2%

Active	Retired
93%	7%

an appropriate alternative. Indeed, small and large claim amounts each displayed distinct model characteristics, which suggest multiple sub-populations underlying the dataset. To illustrate this, Figure 1 displays the empirical cumulative distribution function (CDF) of the claims amount. The left side graph shows the empirical CDF for values below 1,000, while the right side table shows the CDF value for some claims amount to underline how slowly we reach 100% for larger claims. One way of improving the GLM modeling was to transform claims using a log transformation, therefore reducing the spread of the variable to model. Figure 2 illustrates logged claims distribution using a histogram. A histogram is a graphical display of continuous data points and exemplifies the shape and the spread of the underlying distribution. The width of each bar represents the interval of the data points values, and the height represents the count of data points in each interval. Drawbacks of this method include the requirement for specifying the intervals width and lacks smoothing. Kernel estimator is a convenient alternative for illustrating a smooth distribution, and the details related to this approach is described in more depth in Appendix A.

Tables 11 to 16 further display how the severity (average claims) varies across different available variables using two-dimensional analysis. One pitfall of the traditional two-dimensional analysis is that it ignores the interaction between different variables. For example, an increase in severity for retired insured might be attributed to the age factor only which is explained by an increased medical risks, but not necessarily to modified behaviours from retired individuals. This is the case in many regression models which defines the age as a significant explanatory variable, but not the retired status. Another example is the variation of average claim per year, which could be attributed to a shift in the *In Force* distribution, and not to inflation factors which are generally associated to the year variable. Based on the two-dimensional analysis, there appears to be a link between several explanatory variables and the claim amount, which will be confirmed in the regression results shown in the next sections.

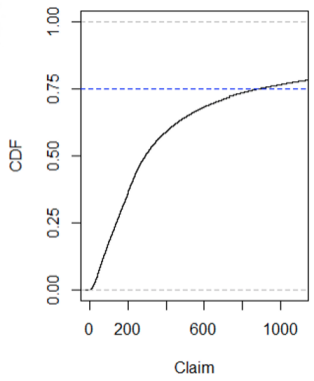
Table 9: Province breakdown

Quebec	Ontario	New Brunswick	Nova Scotia	British Columbia	Alberta	Manitoba	Other
50%	18%	12%	11%	2%	3%	1%	3%

Table 10: Claims Statistics

Min	1st Qu.	Median	Mean	3rd Qu.	Max.
1.30	138.90	301.40	2,373.70	955.20	1,096,391.10

Figure 1: Graph of CDF for claims $< 1{,}000$\$ (left) and table of CDF for claims $\geq 1{,}000$\$ (right)



Claim Amount y	CDF(y)
1,000	76.47%
2,500	86.25%
5,000	91.76%
10,000	95.13%
50,000	99.31%
100,000	99.75%

Figure 2: Histogram of log(Claim)

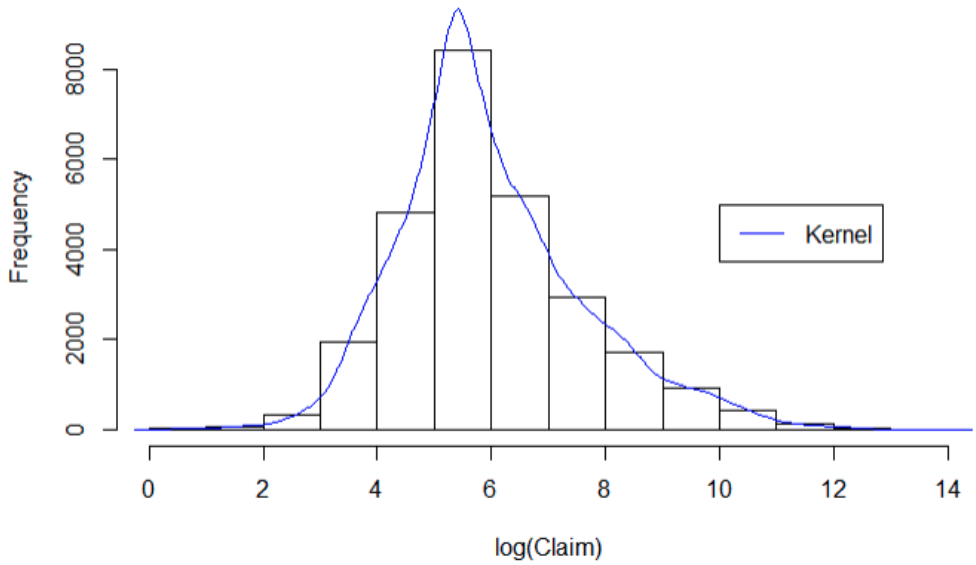


Table 11: Age analysis

Age	Average Claim	Number of Claims
15-19	1,240.18	79
20-29	2,492.36	1,049
30-39	1,223.42	4,315
40-49	1,535.09	5,004
50-59	2,117.70	5,389
60-69	3,060.59	5,731
70-79	4,258.37	3,109
80-89	6,040.50	560
90+	12,193.62	33

Table 12: Gender of Main Insured (Left) and Gender of Claimant (Right) analysis

Gender	Average Claim	Number of Claims
Female	2,319.04	12,066
Male	2,418.03	14,865

Gender	Average Claim	Number of Claims
Female	2,658.67	3,398
Male	3,415.28	2,891

Table 13: Year analysis

Year	Average Claim	Number of Claims
2014	3,327.05	1,720
2015	2,320.56	4,651
2016	2,238.01	7,002
2017	2,426.62	7,636
2018	2,230.63	5,922

Table 14: Company analysis

Company	Average Claim	Number of Claims
Client 1	2,544.55	5,479
Client 2	3,006.48	6,289
Client 3	2,049.48	15,163

Table 15: Salary analysis

Salary	Average Claim	Number of Claims
$10,000 - 19,999$\$	3,010.20	104
$20,000 - 29,999$\$	3,391.43	314
$30,000 - 39,999$\$	3,899.99	656
$40,000 - 49,999$\$	1,849.32	917
$50,000 - 59,999$\$	3,665.22	1,287
$60,000 - 69,999$\$	2,133.46	1,456
$70,000 - 79,999$\$	2,239.74	1,969
$80,000 - 99,999$\$	2,022.44	1,225
$100,000 - 149,999$\$	2,698.00	748
$150,000$\$+	2,699.78	608

Table 16: Status analysis

Status	Average Claim	Number of Claims
Active	2,012.68	21,800
Retired	3,907.42	5,131

3 Generalized Linear Model

This section introduces the theoretical concepts of Generalized Linear Models (GLMs), which are often regarded as the most appropriate models for an *Incidence and Severity* analysis. Indeed, they enable to relate explanatory variables to a response variable with mutliplicative factors which can easily be interpreted and used for actuarial rating. Furthermore, they allow great flexibility for the probability distribution function of the response variable which is in most cases more appropriate than the normal distribution used in Ordinary least squares (OLS) regression. Indeed, a conventional linear regression model is defined as:

$$\mathbb{E}(y|\mathbf{x}) = \mathbf{x}^\top \beta$$

where y is limited to the normal distribution. In a GLM, a link function g relates the expectancy of the response variable to a linear combination of the explanatory variables:

$$g\{\mathbb{E}(y|\mathbf{x})\} = \mathbf{x}^\top \beta$$

where y has a distribution belonging to the exponential family. Note that the vector of explanatory variables $\mathbf{x}$ related to each response variable y is considered fixed, and for this reason, the couple is referred to as $y|\mathbf{x}$. In this report, y refers to the claim amount.

We will also consider the transformation t of a response variable y in a GLM:

$$g\{\mathbb{E}(t(y)|\mathbf{x})\} = \mathbf{x}^\top \beta$$

$$\mathbb{E}(t(y)|\mathbf{x}) = g^{-1}(\mathbf{x}^\top \beta)$$

where $t(y)$ has a distribution belonging to the exponential family and y is the claim amount we are interested in. Using the identity transformation function $t(y) = y$ yields the classic GLM, but a variable transformation appeared extremely useful to obtain an exponential family distribution. Indeed, logged claims displayed a distribution function a lot closer to the normal and the gamma distribution, but this additional step required transformations of the predictions to retrieve the expectation of the original variable of interest. This step significantly reduced the predictive performance of the model as discussed in later sections.

In Section 3.1, we detail the characteristics of the exponential distribution family. We then introduce the regression model of a GLM in Section 3.2 and how the parameters estimation is made in Section 3.3. Section 3.4 outlines how Machine Learning techniques can be used as an alternative method for predictive modeling.

3.1 Exponential Family

A GLM requires that the dependent variable y (or $t(y)$), also referred to as the response variable, follows a probability distribution belonging to the exponential family. To simplify, in the rest of the section we consider that y is the response variable. Below is the corresponding general form of the probability density function:

$$f(y;\theta,\phi) = c(y,\phi) \cdot \exp\left\{\frac{y\theta - a(\theta)}{\phi}\right\}$$

where θ is the canonical parameter, ϕ is the dispersion parameter and $a(\theta)$ defines y's distribution, which is limited to the exponential family (e.g. binomial, normal, poisson, gamma and inverse gaussian). Below are the corresponding closed-form formulas for the expected value and the variance [1]:

$$\mathbb{E}(y) = \dot{a}(\theta),\ Var(y) = \phi \cdot \ddot{a}(\theta) = \phi \cdot V(\mu)\ .$$

From these formulas, we better depict how the variance is related to the expected value. Details related to the calculation of the exponential family parameters above are available in Appendix B.

3.2 Regression Model

In a GLM, the outcomes y_i given $\mathbf{x}_i$, denoted as $y_i|\mathbf{x}_i$ with $i \in \{1, \cdots, n\}$, are assumed independent. For the rest of this document, we consider that we have acces to n data points $y_i|\mathbf{x}_i,\ i \in \{1, \ldots, n\}$. Below is the probability distribution function for each y_i:

$$f(y_i; \theta_i, \phi) = c(y_i, \phi) \cdot \exp\left\{\frac{y_i\theta_i - a(\theta_i)}{\phi}\right\}$$

where f is the density of an exponential family distribution. Furthermore, the expected value of each response variable must be related to a linear component as follows:

$$\mathbb{E}(y_i|\mathbf{x}_i) = \mu_i = g^{-1}(\mathbf{x}_i^\top \beta)$$

where g is referred to as the link function, and satisfies the following conditions: monotonic and differentiable (e.g. identity, log or square root).

One key consideration in the probability distribution function of y_i is that the canonical parameter θ_i is specific to each claim i while the dispersion parameter ϕ is uniform across all risks:

$$\mathbb{E}(y_i|\mathbf{x}_i) = \dot{a}(\theta_i) = \mu_i = g^{-1}(\mathbf{x}_i^\top \beta)$$

$$Var(y) = \phi \cdot \ddot{a}(\theta_i) = \phi \cdot V(\mu_i)\ .$$

Based on these formulas, the expected value of y_i will vary with respect to the vector of attributes $\mathbf{x}_i$, while the variance is related to the combination of the expected value and a fixed parameter ϕ among all risks i. Here is how the variance is linked to the expected value:

$$Var(y) = \phi \cdot V(\mu_i) = \phi \cdot \ddot{a}(\theta_i) = \phi \cdot \frac{d\dot{a}(\theta)}{d\theta} = \phi \cdot \frac{d\mu_i}{d\theta}\ .$$

Since the variance may vary accross data points given its dependance to the mean, we can encounter heteroskedasticity. This aspect of the GLM therefore introduces greater flexibility than the more restricted homoskedastic assumption made in a classic linear regression, in which the distribution is limited to the normal.

In this report, we will often refer to the gamma distribution, which is appropriate for modeling claim amounts

[1] $\dot{a}(\theta)$ denotes the first derivative of $a(\theta)$ and $\ddot{a}(\theta)$ denotes the second derivative of $a(\theta)$

given that values are restricted to positive real numbers. Below is the distribution function of the gamma using the (α, μ) parametrization ((Jong & Heller, 2008), Section 2.7 Chi-square and gamma, p.27):

$$f(y; \alpha, \mu) = \frac{y^{-1}}{\Gamma(\alpha)} \left(\frac{y\alpha}{\mu} \right)^{\alpha} e^{-y\alpha/\mu} \tag{1}$$

where

$$\mathbb{E}(y) = \mu$$

$$Var(y) = \mu^2/\alpha \ .$$

Here is how equation (1) can be re-written:

$$f(y; \alpha, \mu) = \frac{y^{\alpha-1}\alpha^{\alpha}}{\Gamma(\alpha)} \cdot \exp\left(\frac{-y/\mu - \ln(\mu)}{1/\alpha} \right) \ .$$

This way, we obtain the corresponding component of the gamma distribution for the canonical parameter θ and the dispersion parameter ϕ:

$$c(y; \phi) = \frac{y^{\alpha-1}\alpha^{\alpha}}{\Gamma(\alpha)}$$

$$\theta = -\frac{1}{\mu} \quad , \quad \phi = \frac{1}{\alpha}$$

$$a(\theta) = \ln(\mu) = \ln(-\frac{1}{\theta}) = -\ln(-\theta)$$

$$\dot{a}(\theta) = -1 \cdot \frac{1}{-\theta} \cdot (-1) = -1/\theta = \mu \ .$$

Details related to the calculations above for the gamma distribution are available in Appendix C.

3.3 Parameters Estimation

The value of the parameters β_j and ϕ are derived with statistical programs like R using Maximum Likelihood Estimation (MLE) and yield the estimations $\hat{\beta}_j$ and $\hat{\phi}_i$. Below is a reminder of the likelihood function $\mathcal{L}$ for the observed data sample:

$$\mathcal{L}(\beta, \phi, \mathbf{y}) = \prod_{i=1}^{n} f(y_i; \beta, \phi) \ .$$

The likelihood function can be expressed as a product of the univariate density function given the independence assumption. The log-likelihood is simply the log of the likelihood function, which converts the multiplication component in the formula above to a summation:

$$l(\beta, \phi, \mathbf{y}) = \log(\mathcal{L}(\beta, \phi, \mathbf{y})) = \log(\prod_{i=1}^{n} f(y_i; \beta, \phi)) = \sum_{i=1}^{n} \log(f(y_i; \beta, \phi)) \ .$$

This change is useful since the gradient method is used to maximize the formula above, and the use of a summation is more convenient for this purpose. Given that the log function is monotonically increasing, maximizing the likelihood is also equivalent to maximizing the log-likelihood. Maximizing the log-likelihood

with respect to parameters β_j and ϕ implies solving the following formulas:

$$\frac{\partial l}{\partial \beta_1} = 0,\ \frac{\partial l}{\partial \beta_2} = 0, \ldots,\ \frac{\partial l}{\partial \beta_p} = 0 \tag{2}$$

$$\frac{\partial l}{\partial \phi} = 0\ . \tag{3}$$

The equations above reach a closed-form with the use of the identity link $g(y) = y$ and the normal distribution, and is equivalent to the Ordinary Least Squares (OLS) solution for linear regression. In other cases, the solution is not as direct and requires an iterative processes to estimate (2) and (3). The Newton-Raphson iteration process is one method which consists of a quadratic approximation of the log-likelihood function using its first and second derivative. Below is the resulting iterative formula for approximating β, assuming it is of one dimension:

$$\beta^{(m+1)} = \beta^{(m)} - \frac{\dot{l}(\beta^{(m)})}{\ddot{l}(\beta^{(m)})} \tag{4}$$

where m refers to the $m - th$ iteration of the iterative process, which is known to converge rapidly ((Jong & Heller, 2008), Section 5.5 Maximum likelihood estimation, p.69). Fisher scoring is an alternative method which replaces the second derivative term $\ddot{l}(\beta^{(m)})$ with its expectation $\mathbb{E}(\ddot{l}(\beta^{(m)}))$ in formula (4). Below is the resulting estimation using this method:

$$\beta^{(m+1)} = (X^\top W X)^{-1} X^\top W \{X\beta^{(m)} + G(y - \mu)\} \tag{5}$$

where G and W are the matrices with diagonal element $\dot{g}(\mu_i) = \frac{\partial x_i^\top \beta}{\partial \mu_i}$ and $\left(\dot{g}(\mu_i)^2 V(\mu_i)\right)^{-1}$ respectively.

For coefficients β_j, Fisher scoring is used and is often equivalent to the Newton-Raphson iterative process((Jong & Heller, 2008), Section 5.5 Maximum likelihood estimation, p.69). The details related to this method is outside of the scope of this report, and is used in R's built-in functions for computing the components of a GLM model.

3.4 Machine Learning

Commonly known problems and restrictions with a GLM are as follows ((Mark Goldburd & Guller, 2020), Section 10 Variations on the Generalized Linear Model, p.93):

- Restriction of the response variable's distribution to the exponential family
- Restriction to a linear function of the predictors $g(\mu) = x^\top \beta$
- The dispersion parameter ϕ of the exponential family is held constant across risks
- Unstable coefficients of regression in the presence of highly correlated variables
- A coefficient linked to a specific feature is given full credibility without regard to the number of data points underlying each feature value

Many items listed above can be addressed using modern Machine Learning techniques, which are not commonly seen among actuarial methods in the Life & Health insurance field. Nevertheless, most of these issues are not of particular concern in this study, but other challenges like the limitation to the exponential family

will be addressed with the use of variable transformation.

Machine Learning is the field that studies algorithms which learn from previous experience without needing explicit programming for improvement. It has attracted much attention in the last years due to computers increased performance, remarkable research advancements and impressive achievements when used in different types of businesses like fraud detection, image recognition and natural language processing.

There are two main types of learning algorithms. Supervised learning is the task of predicting a specific outcome linked to an input example, given that an algorithm was previously trained using a training dataset containing multiple pairs of input x and output y. This type of data is more generally referred to as a labelled training data. Unsupervised learning will instead focus on data inference and seek for patterns and clusters, without existence of labelled responses. The former would be the task of interest for modeling claim occurrence.

This report specifically focuses on the use of Generalized Linear Models given the amount of literature and the proven adequacy of this method for modeling claims severity. That being said, additional analysis should be performed to validate if Machine Learning methods like Random Forest and Neural Networks would also be appropriate, and gain insights using these alternative methods.

4 Model Building

The purpose of a GLM model in this context is to predict the claim amount for an incurred claim, given different variables linked to the insured coverage. The GLM is extremely useful since it allows the target prediction variable to follow a different probability distribution than the normal as in a normal linear regression. Available explanatory variables include features such as the age, salary or the province of residence, which are listed in table 4 from Section 2. The model for the claim amount is referred to as a severity model. An incidence model, instead, predicts a claim occurrence by modeling the probability of a claim taking place, but this remains outside of the scope of this report. The combination of an incidence and severity model constitutes a full model for claim modeling.

In Section 3, we have defined a GLM as:

$$y_i|\mathbf{x}_i \sim f(\theta_i, \phi)$$

where f is a probability distribution from the exponential family, and

$$g(\mu_i) = \mathbf{x}_i^\top \beta = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}$$

where μ_i is the expected value of the i-th claim amount y_i, $\mathbf{x}_i = (1,\ x_{i1},\ x_{i2}, \cdots,\ x_{ip})$ and $\beta = (\beta_0,\ \beta_1,\ \beta_2, \cdots,\ \beta_p)$. A model predicts the value of y_i with $\hat{y}_i$, defined as

$$\hat{y}_i = \hat{\mu}_i = g^{-1}(\mathbf{x}_i^\top \hat{\beta})\ .$$

This section focuses on the model selection steps involved in building a GLM for the severity in group travel insurance. These steps consist of defining the claims distribution f, the link function g and choosing the explanatory variables x, each described in Sections 4.1, 4.2 and 4.3. In this report, we also address the choice of a transformation function $t(y)$ so that the distribution function f from the exponential family is a better fit to $t(y)$ than to y. Such transformations are usually not recommended since they tend to reduce the interpretability of the regression coefficients. Furthermore, statistical measures resulting from the regression on the transformed variable $t(y)$ are not relevant for the original variable of interest y (Feng, 2014). The choice of transformation function t is not considered as a model building step since it implies comparing models with different target variables, which prohibits the use of common statistical criterion for model comparison such as the AIC (see Section 5 for more details). We will not address this issue given that the predictive performance of the model involving a transformation proved disappointing.

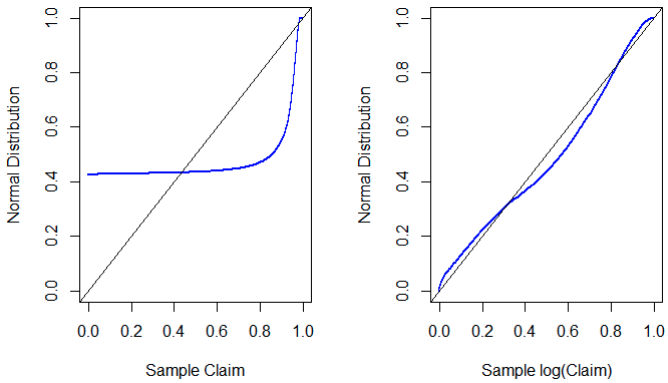
A key consideration when building a model is to prevent over-fitting to the data used to train the model. A model obtaining a perfect fit to a dataset might appear as a good candidate, but if it is unable to predict properly the value of a new and unseen data, then it indicates poor generalization. Therefore, it is crucial to select a model that trades off between high complexity with low bias and high variance, versus simplicity with high bias and low variance. This concept is defined as the Bias-Variance trade-off and optimally targets a combination of low bias and low variance. A simple training and test dataset split is used to build a model in the first place with the train dataset, then compute the generalization error on new unseen data with the test dataset. A split of 75% for the training dataset, and the remaining 25% for testing the model that was chosen.

4.1 Probability Distribution function f

Common choices of exponential probability distribution for claims severity are the gamma and inverse gaussian. Normal distribution is usually not a good candidate because the variable of interest is non-negative and skewed to the right, but it remains a good consideration in model comparison due to its simplicity.

Figures 3, 4 and 5 illustrate how the the claims y_i's distribution compares to the normal, gamma and inverse gaussian distributions respectively, without regard to the vector of explanatory variables $\mathbf{x}_i$. The parameters of the fitted probability distribution were derived using R's Maximum Likelihood (MLE) built-in functions. For each data point of the sample, the percentile rank is compared to the parametric distribution percentile rank. The y-axis consists of the parametric distribution percentile rank while the x-axis is the sample percentile rank. The straight line ($y = x$) would be the resulting pp-plot with perfect correspondence between the parametric distribution and the sample rank percentile. Parametric distributions with pp-plots close to the linear function are therefore considered a very good fit. Based on the illustrations, using a gamma distribution appears to be the most appropriate fit. One pitfall of comparing the sample's distribution to a probability distribution function is that we ignore the contribution of the explanatory variables in the model, and therefore assume uniform exponential parameters among all risks. In other words, a GLM yields different distribution parameters (θ_i, ϕ) for each data points y_i depending on the values of the vector $\mathbf{x}_i$, while the MLE estimation used in the pp-plot yields the same parameters (θ, ϕ) for all y_i's. Therefore, the method described here is a gross approximation for choosing the appropriate distribution, and the statistical results for each GLM models will more accurately guide the final choice.

Figure 3: PP-plot of claims with Normal Distribution



Nevertheless, pp-plot with MLE estimation remains a reasonable method for assessing claims distribution. Based on the illustrations, the most appropriate fit appears between the logged claims and Gamma distribution in Figure 4. In this report, we will only consider the Gamma distribution for the candidate models presented in the next sections.

The linear component of the GLM function suggests that all claims y_i with the same vector of explanatory variables $\mathbf{x}_i$ also share the same expected value μ_i. Therefore, it is assumed that every y_i should follow the

Figure 4: PP-plot of claims with Gamma Distribution

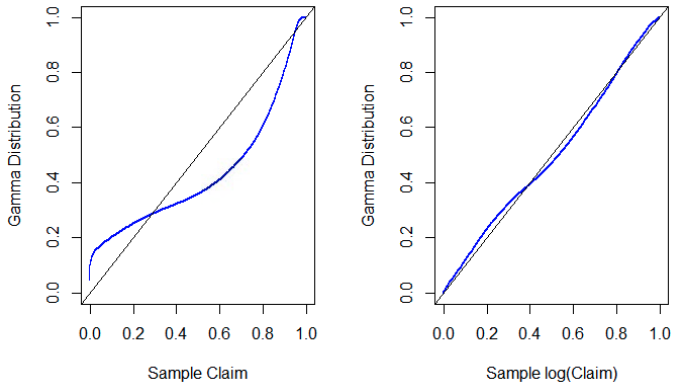
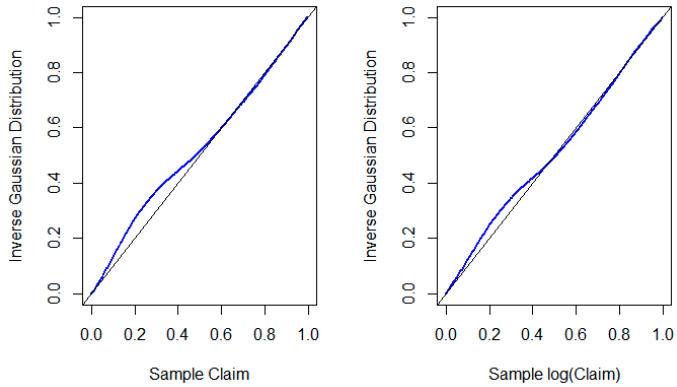


Figure 5: PP-plot of claims with Inverse Gaussian Distribution



same parametric exponential family, but with different parametric values (θ_i, ϕ) for each possible vector $\mathbf{x}_i$. As an example, Figure 6 displays how logged claims vary for each participating partner, also referred to as client, and for each age group, which are expected to be key variables in the severity model. As shown, the shape of the distribution doesn't seem to change significantly with respect to each variable. Again, this method is extremely approximative and doesn't account for the interaction with other variables (which should be held constant), but it still underlines that the claims distribution is fairly constant across different groups of risk and suggests that a GLM model could be appropriate in this context.

4.2 Link function g

In a GLM, $g(\mu_i)$ is linked to x_{ij} through a linear function:

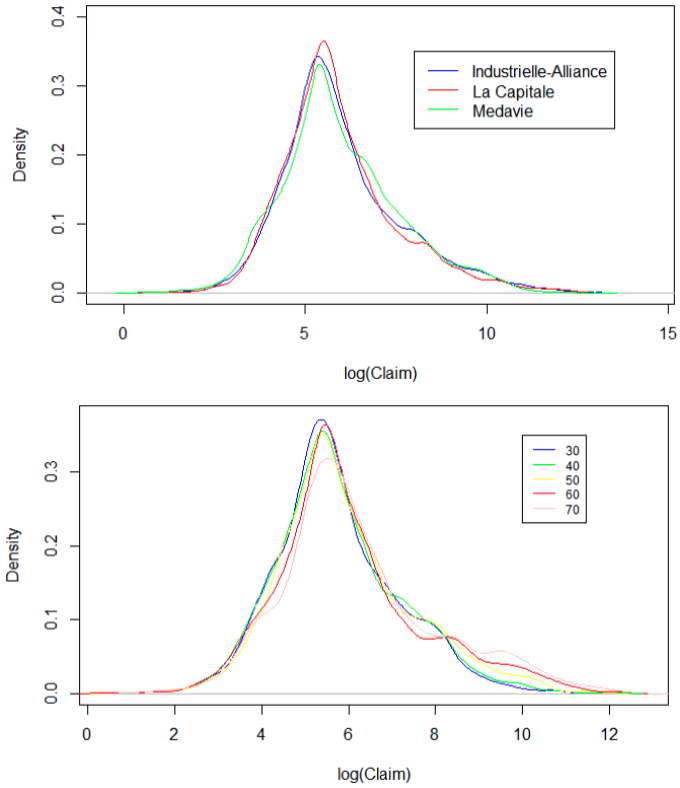
$$g(\mu_i) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_j x_{ij} + \cdots + \beta_p x_{ip} \ .$$

An identity link $(g(x) = x)$, as in a classical linear model, would yield the following relation:

$$g(\mu_i) = \mu_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_j x_{ij} + \cdots + \beta_p x_{ip} \ .$$

With an identity link function, coefficients β_j are additive. For example, let's suppose β_{age} is the coefficient for the age variable. Therefore, for each additional year of age, the expected claim would increase with an

Figure 6: Density split per client (Up) and per Age group (Down)



additive factor of β_{age}\$. An alternative approach is to use the log-link function:

$$g(\mu_i) = \log(\mu_i) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$$

$$\mu_i = e^{\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}}$$

$$\mu_i = e^{\beta_0} e^{\beta_1 x_{i1}} \cdots e^{\beta_p x_{ip}} \ .$$

With a log-link function, β_{age} yields a multiplicative factor of $e^{\beta_{\text{age}}}$. This form improves significantly the interpretation of the coefficients, especially when insurance rating is the task of interest. As an example, let's consider the regression coefficient related to the age of the insured $\beta_{\text{age}} = 0.03$. Given that the coefficient is non-negative, it suggests that the risk of larger claims increases with age. The related multiplicative factor r_{age} is defined as

$$r_{\text{age}} = e^{\beta_{\text{age}}} \approx 1.03045 = 103.03045\%$$

and implies that for each additional year of age, a claim is expected to increase of +3.03045%. These factors are fairly easy to implement in an actuarial pricing process and can be compared to historical ratings which are also based on multiplicative factors.

Here, we should not confuse the log-link with the log transformation. The log transformation is different and can be expressed as below with an identity link:

$$\mu_i = E(log(y_i)|\mathbf{x}_i) \neq log(E(y_i|\mathbf{x}_i)) \ .$$

Furthermore, it is crucial to understand how the expected value μ_i varies according to the continuous explanatory variables x_{ij}. A simple way to illustrate this is to graph the average claims, as an approximation for μ, with respect to different values of a specific continuous explanatory and check for linearity. Such continuous explanatory variables include the age and the salary. The step of averaging the claim amount mistakenly ignores the interaction with other explanatory variables (which should be held constant), but this is a simplified method which prevents lack of data and high volatility per risk class. Transformation g is further applied to the explanatory variable to validate if it yields enhanced linearity. As suggested in the previous section, the log link is a very good candidate for interpretability purpose.

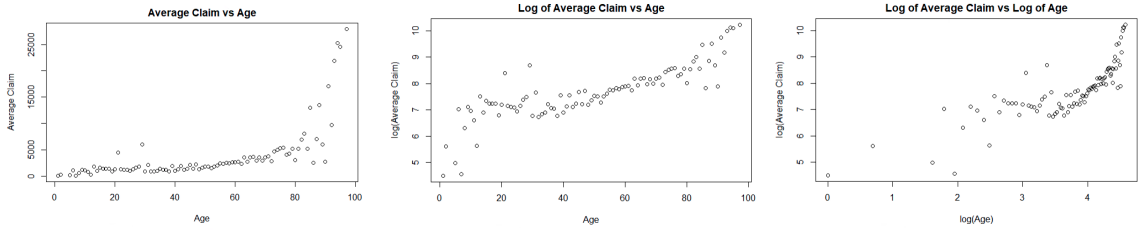
Here is how the average claim was computed for each explanatory variable x:

$$\bar{y}_x = \frac{1}{n_x} \sum_{i=1}^{n} y_i \cdot \mathbb{1}\{x_i = x\}$$

where n_x is the number of claims y_i for which the explanatory variable x_i equals the specific value x. For example, using the age as the explanatory variable of interest, the resulting y value for $x_{\text{age}} = 35$ years equals the average of all claims for which the age of the claimant is 35.

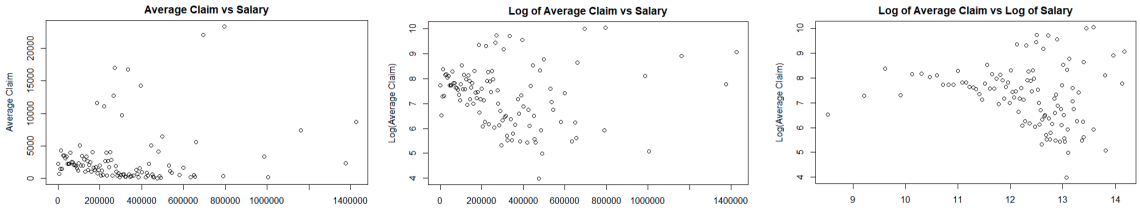
Figures 7 and 8 illustrate how the average claim varies with respect to the age and the salary respectively, depending on how the log transformation is applied. The first graph from the set shows the relationship without any transformation, the second applies the transformation on the average claim, and the third applies the transformation on both the average claim and the explanatory variable. Ignoring outliers, both sets of graphs suggest that the log of the average claim is linearly linked to the salary and the age. Indeed, the second graph from each set displays a linear trend, which is not improved when logging the explanatory variable. Outliers around the age of 30 in Figure 7 are caused by few very large claims, and are more related to exceptional events than to the explanatory variables, while the sudden increase in older age is related to higher medical risk. In Figure 8, outliers are more frequent given that the salary is not available for every insured, which reduces the number of claims for each data point in the graph and therefore yields higher variability. Nevertheless, the decreasing trend with respect to the salary appears as linear in the second graph and doesn't visibly improve in the first or in the third graph. We conclude that no transformation of these two continuous explanatory variables is required in the model, if they are considered significant in the regression model.

Figure 7: Claim vs Age with log transformation



In the previous section, we have also suggested that modeling the logged claims could improve fit to the gamma distribution. Therefore, we can repeat the same procedure described above, and replace the target

Figure 8: Claim vs Salary with log transformation



variable to $\log(y)$, and analyse how the log of average logged claims $\log(\overline{\log(y)})$ behaves with respect to the age and salary. Illustrations are available in Figures 9 and 10. Here, linearity appears when the average claims are logged or not, which might suggest the identity link as a potential candidate for a GLM with the target variable $log(y)$. A polynomial transformation of the age variable could also seem like a reasonable candidate based on the first graph of Figure 9, but we will prefer the identity link for simplicity.

Figure 9: log(Claim) vs Age with log transformation

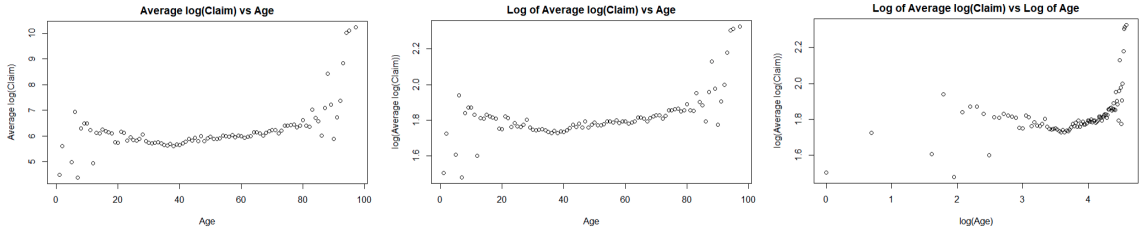
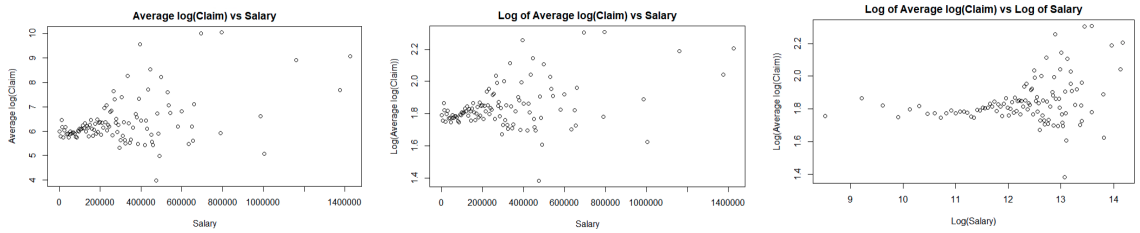


Figure 10: log(Claim) vs Salary with log transformation



4.3 Explanatory Variables x

In group travel insurance, only partial information can be used in a predictive model, since most details related to a claim are not known at the moment of pricing. For example, a group insurance rating cannot account for the destination of the trip or the medical diagnosis related to the claim, because this information is only available once the insured event has occurred. Therefore, the model should only account for explanatory variables which are known when a group insurance policy is issued. The full list of these fields is available in table 4 from Section 2, and includes real-valued, binary and and categorical fields.

4.3.1 Binary and categorical variables

Binary and categorical variables are defined as factors that can take one of the value from a predefined set of values. In the case of a binary field, there are only two possible values while for a categorical field, there are

more than 2 possible values. For example, the gender of an insured is a binary field and can be either female or male, while the protection type is a categorical field and can be either couple, family, single or single parent.

The way a GLM treats categorical variables is that it defines a baseline factor, and then computes a specific coefficient for all other factor values. For example, let's say that the baseline factor for the province variable is Quebec. The GLM will then set the Quebec variable $x_{i,\text{quebec}}$ to 1 for all claims y_i for which the insured is a resident of Quebec, and 0 otherwise. Then, another variable is created for each other province ($x_{i,\text{ontario}}$, $x_{i,\text{british}-\text{columbia}}$, etc...), and the values of 0 and 1 are set equivalently as for the baseline factor. Suppose there are n_{prov} different provinces, then $n_{prov} - 1$ different province coefficients ($\hat{\beta}_{\text{ontario}}$, $\hat{\beta}_{\text{british}-\text{columbia}}$, etc...) will be defined in the GLM, and the baseline factor's estimate $\hat{\beta}_{\text{quebec}}$ will be included in the intercept value $\hat{\beta}_0$.

Selecting the appropriate baseline factor is a key step in the selection of explanatory variables. Indeed, the significance of the coefficients of regression, which can be measured using the p-value as detailed in Section 5, is sensitive to the baseline factor value, and one should select a factor which is considered as a good value of reference. Furthermore, selecting a baseline factor with a larger number of claims should improve the model's robustness. For example, a reasonable baseline factor for the province of residence would be Quebec, since it accounts for the largest number of claims and we are interested in knowing how other province's claim severity compares to this province.

4.3.2 Real-valued variables

Real-valued variables include the age, the salary and the year. In some cases, grouping the real-valued variables into buckets and treating them as categorical variables can improve the predictive performance of a model, but it can also lead to important over-fitting. Furthermore, Figures 7 to 10 have shown trends which appear closer to linearity, and justifies the use of real-valued variables instead of buckets. We have also preferred to not apply any transformation to the real-valued variables since this step doesn't improve the linearity trend with respect to the severity variable as discussed in Section 4.2.

4.3.3 Forward and backward selection

To select which explanatory variables to include in the model, both the forward and the backward selection methods were used. The forward method consists of starting with no variable at all, then adding the variable which increases the most the chosen fitting measures. These measures are described in Section 5 and consist in this report to the Deviance and the AIC. The backward method instead starts will all available variables included in the model, and the least significant is removed at each iteration until all variables are significant. The measure of significance is the p-value, and is described in Section 5. Both methods were tested to make sure they lead to the same selection of explanatory variables, which was possible due to the few numbers of available fields. Depending on the choice of the response variable transformation $t(y)$ and the distribution function $f(y)$, this process has yield different sets of explanatory variables, and the results are detailed in Section 7.

5 Fitting measures and assessment

To discriminate models and select variables with care, the use of fitting measures is essential. This section presents the most useful ones for this analysis. As a reference, Figure 11 displays an output from R's GLM summary function.

Figure 11: GLM output

```
Call:
glm(formula =  Claim ~ Variable 1 + Variable 2 + Variable 3 + Variable 4 +
               Variable 5 + Variable 6              = Gamma(link = "log"),
    data = train_sub)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.9806  -0.8320  -0.2643   0.3469   1.7793

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                  5.495e+00  2.678e-02 205.214  < 2e-16 ***
Variable 1                   3.334e-01  3.511e-02   9.497  < 2e-16 ***
Variable 2                   1.932e-01  1.751e-02  11.032  < 2e-16 ***
Variable 3                   1.570e-02  5.576e-03   2.816 0.004872 **
Variable 4                  -3.400e-02  1.457e-02  -2.333 0.019637 *
Variable 5                   4.541e-02  2.225e-02   2.041 0.041312 *
Variable 6                   7.479e-07  1.928e-07   3.879 0.000105 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Gamma family taken to be 0.670264)

    Null deviance: 11871  on 15293  degrees of freedom
Residual deviance: 11700  on 15287  degrees of freedom
AIC: 201870

Number of Fisher Scoring iterations: 6
```

The Deviance Δ and the AIC are two measures which were used both for the forward/backward explanatory variables selection step, and for comparing candidate models with different distribution function g and link function g. These measures are introduced in Sections 5.1 and 5.2. The final model was then validated by comparing its deviance residuals, also referred to as residuals, to a normal distribution for fitting assessment, as explained in Section 5.3. Section 5.4 introduces the p-value, which is the measure of significance of the explanatory variables, which was used in the backward selection method described in Section 4.

5.1 Deviance Δ

Deviance Δ is a measure of how far a model is from a perfect fit. A perfect fit is achieved using a saturated model, and includes n regression coefficients for a sample dataset of size n. In the saturated model, every single data point y_i is equal to its fitted value, hence $\hat{\mu}_i = y_i$. The reason for a perfect fit is that each coefficient $\hat{\beta}_{ii}$ equals the estimation $\hat{\mu}_i = y_i$ $(\hat{\beta}_{ij} = 0,\ i \neq j)$, although each coefficient doesn't have any predictive relevance. Furthermore, such model will hardly predict new unseen data accurately given that the model is over-fitted to the data that it was trained on. Such a model is a good example of a lack of Bias-Variance trade-off, with extremely high variance and null bias. Nevertheless, it yields maximum log-likelihood given a distribution function f and link function g. For the same combination of functions f and g, the minimum log-likelihood model is one which includes no coefficients at all, with a resulting prediction $\hat{\mu}_i = \hat{\beta}_0$ (intercept) for all i. This is referred to as the null model.

Deviance for a specific model is defined as ((Jong & Heller, 2008), Section 6.1.2 Deviance, p.63):

$$\Delta = 2 \cdot (l_{saturated} - l_{model}) \equiv 2 \cdot (\check{l} - \hat{l}) \tag{6}$$

where $\check{l} \geq \hat{l} \geq l_{null} \equiv \bar{l}$. Indeed, $\check{l}$ is the likelihood of the saturated model, which is the maximum likelihood that can be observed, while $\bar{l}$ is the likelihood of the null model, which is, alternatively, the minimum likelihood to be observed. The $\hat{l}$ is the likelihood of the model we are testing, which we are trying to maximize. The likelihood is not a reasonable standalone fitting measure since its value doesn't reveal much about the predictive performance of the model, which is better assessed when compared to the likelihood of the saturated model. The log-likelihood l formula was defined in section 3.3.2, and it can be expressed more simply in the deviance equation knowing that the distribution function f is from an exponential family ((Jong & Heller, 2008), Section 5.7 Assessing fits and the deviance, p.72):

$$\Delta = \sum_{i=1}^{n} 2 \cdot \frac{y_i(\check{\theta}_i - \hat{\theta}_i - a(\check{\theta}_i) + a(\hat{\theta}_i))}{\phi} \equiv \sum_{i=1}^{n} \delta_i^2$$

where δ_i^2, the i-th deviance residual, is the individual contribution of the i-th estimation to the global deviance Δ. The sign of δ_i is taken to be the sign of $y_i - \hat{\mu}_i$ ((Jong & Heller, 2008), Section 5.9 Residuals, p.78). The intuition of this equation is to measure how far the likelihood of a tested model is from the likelihood of a model with perfect prediction.

Deviance Δ is known to be asymptotically equivalent to a Chi-Square χ^2_{n-p} distribution as $n \to \infty$ with expected value $n - p$ ((Jong & Heller, 2008), Section 5.7 Assessing fits and the deviance, p.72). Conditions required for this to hold is if the fitted model is well assessed, ϕ is known and n is large enough. Given that

$$\mathbb{E}(\chi^2_{n-p}) = n - p \ ,$$

where $n-p$ is the number of degrees of freedom, we expect the following condition when a good fit is assessed:

$$\frac{\Delta}{n-p} < 1 \ .$$

Otherwise, the model has a poor fit. If ϕ is unknown and is estimated instead, then the χ^2 distribution hypothesis is not reliable anymore and one shouldn't consider this statistical measure as a criterion for fitting assessment. Here, the deviance should only be used to compare two nested models, meaning when a smaller model (in number of coefficients) is a special case of a larger model to which it is being compared ((Jong & Heller, 2008), Section 5.7 Assessing fits and the deviance, p.72). Indeed, it is proven that the difference of deviance for a sequence of nested models is approximately χ^2, but the χ^2 approximation for the standalone deviance should not be relied on (P. McCullagh, 1989). Therefore, the deviance is used more as a screening device for selecting explanatory variables between nested models, rather than as a precise significance measure.

R will output two deviance measures: the Null and the Residual deviance. They are respectively the equivalent deviance for the model including only the intercept β_0 and all explanatory variables β_j specified in the regression model. In order for the explanatory variables to be considered as significant in the model, one should expect a reduction in the residuals deviance when compared to the null deviance value.

5.2 Akaike's Information Criterion (AIC)

The AIC is used as a baseline to compare different models with each other:

$$AIC \equiv -2l + 2p$$

where l is the maximum log-likelihood and p is the number of parameters of a model.

The AIC formula is widely used because it balances between low bias by seeking a higher log-likelihood value, and low variance by penalizing models that are too complex with higher p values. This statistical property is commonly referred to as the Bias-Variance tradeoff. This prevents over-fitting the data that was used to select the model and enabling it to generalize to new unobserved data with a reasonable complexity structure. Intuitively, what the formula of the AIC tells is that if a more complex model, for example a model adding a new coefficient of regression, has an increased AIC value, then the increase of likelihood offsets the increased complexity from the added coefficient. It is also interesting to mention that the leave-one-out cross-validation method is asymptotically equivalent to using the AIC (Stone, 1977), which justifies even more the use of this simplistic fitting measure.

An alternative measure to the AIC is the Bayesian Information Criterion (BIC):

$$BIC \equiv -2l + p\log(n)$$

The formula above is very similar to the AIC, but penalizes more heavily for higher number of variables p due to the $\log(n)$ component, and therefore tends to select simpler models. This measure is known to be model consistent, meaning that it indicates which is the real model to use as $n \to \infty$. Therefore, this measure should be used when we are trying to reach the real model. In this report, we will prioritize the use of the AIC given that few explanatory variables are available, that actuarial judgement will be used for selecting them and that predictive performance will also be a key component in selecting the model.

It is important to note that only models with the same sample of dependent variables y_i should be compared to each other using this measure. Indeed, the log-likelihood will change significantly if a transformation is applied to the y_i's or if a different sample is used. In this case, the AIC for two models with the same structure and different sample data points cannot be compared. Furthermore, the standalone scale of the AIC is also useless, since it is dependent on the number of data points, and should only be compared with respect to alternative models using different probability functions f, link functions g or explanatory variables x.

5.3 Residuals

Residuals in linear models with OLS regression are defined as

$$\hat{\epsilon}_i = y_i - \hat{y}_i$$

, with $\hat{y}_i = \hat{\mu}_i$ and are proven to be normally distributed. Unfortunately, such a simplistic definition of residuals doesn't hold for a GLM, and alternative residuals measures are available for assessing the fit of a

model. The most common one is the deviance residual δ_i^2 as referred to in subsection 5.1, where

$$\delta_i = sign(y_i - \hat{\mu}_i) \cdot \sqrt{\delta_i^2}$$

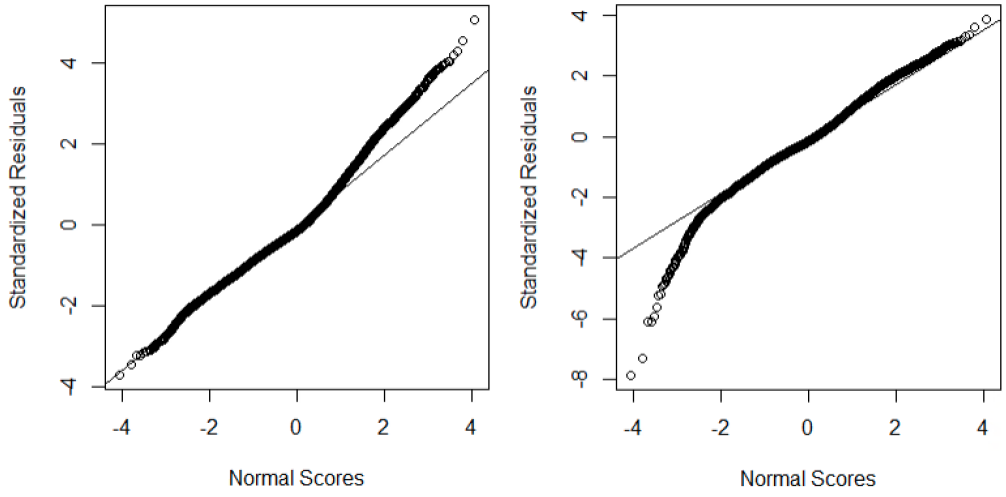
(for the linear model, $\delta_i = \hat{\epsilon}_i$). It was stated that the deviance

$$\Delta = \sum_{i=1}^{n} \delta_i^2$$

is asymptotically equivalent to a $\sim \chi^2_{n-p}$, and from that, it is generally recognized that $\delta_i \sim$ Normal distribution (Portugués, 2020). Using the standardized deviance residual, we can check if the deviance residuals follow a $N(0,1)$ distribution using a qq-plot. A qq-plot compares the quantile of a sample dataset with the quantile of a theoretical distribution function, which in this case is the standardized Normal distribution. In order for the Normal assumption to hold, the response distribution f must be well chosen and the sample must be reasonably large. Therefore, plotting the deviance residuals distribution can help discriminate response distributions.

Figure 12 illustrates as an example the fit of standardized residuals to a normal distribution $N(0,1)$ with a qq-plot for two models, one which uses the Normal distribution and the other the Gamma distribution for claim severity. This graph suggests a lack of fit on the right tail side with a Normal distribution, and a lack of fit on the left tail side with a Gamma distribution.

Figure 12: qq-plot of Standardized Residuals and Normal Scores with a Normal (Left) and Gamma (Right) distribution



5.4 p-value

When selecting explanatory variables in a model, it is common to refer to its significance using p-values. The corresponding column for the p-value in the summary report is labelled $P(>|t|)$. The associated t value

refers to a test statistic. If $\sigma_{\hat{\beta}}$ is known, it is the z-score and is equal to

$$z = \frac{\hat{\beta} - \beta_0}{\sigma_{\hat{\beta}}}$$

which is asymptotically equivalent to a $N(0, 1)$. If $\sigma_{\hat{\beta}}$ is unknown and the dispersion parameter ϕ must be estimated, it can be replaced with $\hat{\sigma}_{\hat{\beta}}$ and we obtain the alternative measure:

$$t = \frac{\hat{\beta} - \beta_0}{\hat{\sigma}_{\hat{\beta}}}$$

which is asymptotically equivalent to a *Student t*. Component β_0 in the formulas above is a known constant, and is needed to test the null hypothesis $H_0 : \beta = \beta_0$. It should not be confused with the intercept of a model. Most statistical package report the test statistic for $\beta_0 = 0$, and therefore, we obtain:

$$z = \frac{\hat{\beta}}{\sigma_{\hat{\beta}}},\ t = \frac{\hat{\beta}}{\hat{\sigma}_{\hat{\beta}}}$$

which is simply the ratio of the first two columns labelled *Estimate* and *Std. Error*. The interpretation of the p-value corresponds to the the probability of observing a test statistic at least as extreme given that the null hypothesis $H_0 : \beta = \beta_0 = 0$ holds. If the p-value is small enough ($< 5\%$), then we reject the null hypothesis, conclude the alternative hypothesis $H_1 : \beta \neq 0$ and include β in the model.

6 Predictive performance

In this report, the goal is to compare different models' predictive performance for the severity in travel insurance claims while retaining a reliable model that is statistically sound. The AIC and the deviance Δ described in Section 5 are strong statistical measures for comparing models with the same data points y_i. Therefore, comparing the predictive performance of a GLM with the claim amount as the response variable with another GLM with the logged claim amount as the response variable cannot rely its comparison on the basis of the statistical measures described in the previous section. In this section, we will discuss measures which quantify the predictive performance of models, regardless of the response variable feeding the model. Indeed, all models can be used to predict the variable of interest. Measures presented in this section are the Root Mean Squared Error (RMSE), the actual versus predicted plot and the actual versus predicted ratio, each described in Sections 6.1, 6.2 and 6.3.

One issue with the predictive models discussed in the results from Section 7 is that the number of available explanatory variables in group travel insurance are very limited. Available information includes details related to a group of insured with active or retired employees, but not specifically to a group of people who are planning on doing a trip as in individual travel insurance. Therefore, there are no information related to the trip itself. Furthermore, no medical underwriting is made in group insurance, given that it is presumed that a group of active employees are known to have lower medical risk. With these limitations, the model derived is not expected to have good predictive performance for each claim. That being said, the model can still have very good global predicting performance when observing how it predicts the average claim for a specific risk class.

6.1 Root Mean Squared Error (RMSE)

Predictive performance is often quantified using the deviation of the predictions $\hat{y}$ from the actual claims y. Two common measures are the Mean Absolute Error (MAE) and the Root Mean Squared Error (RMSE), where

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

and

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}\ .$$

The RMSE has a tendency of attributing more weight to large errors due to the squared component. Therefore, we will prefer the RMSE over the MAE for predictive assessment. In a GLM model, the predicted value $\hat{y}_i$ is replaced with $\hat{\mu}_i$, as defined in Section 4.

6.2 Actual vs Predicted Plot

Another way of assessing the predictive performance of a model, regardless of the model specifications, is to compare actual values y_i with their predicted values $\hat{\mu}_i$. In an idealistic scenario, we want $y_i \approx \hat{\mu}_i$. Therefore, a graph illustrating actual values on the y-axis for each corresponding predicted value on the x-axis should

appear close to a straight line with equation $y = x$ if the prediction is good. Given that a model with a large number n of data points $y_i,\ i \in 1, \ldots, n$ and few explanatory variables x_{ij} as input might lead to unreasonable individual prediction for each data point i, it is recommended to aggregate the actual and predicted average claims in buckets. For this, we order the dataset in increasing order of predicted values, and split them in 100 buckets, each referring to a specific percentile rank. Then, the average of actual claims and the average of predicted claims are computed for each bucket, and act as the input values for the plot of actual versus predicted values.

6.3 Actual vs Predicted Ratio

Actuaries often asses the predictive performance of a model using Actual (A) to Predicted (P) ratios. The global A/P ratio is equal to the total claim amount divided by the total predicted claim amount. We could equivalently compute the ratio of the average claim to the average predicted claim. This measure doesn't reveal much about the error made on each prediction, and only enlightens how the model performs globally at predicting the total claim amount (or average claim amount). For example, a model could use the average total claim as a prediction, i.e. $\hat{y}_i = \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \ \forall\ i$ and result in a perfect A/P ratio, but also yield an extremely large RMSE due to high errors on each prediction. Nevertheless, such ratios can be useful when it is computed on a test set and is split per risk class based on the values of explanatory variables, and helps to evaluate if a model generalizes well to new unseen data points.

7 Results

Concepts and methods introduced in the previous sections are used here for testing different models and data transformations to derive an appropriate statistical model for the prediction of claims in group travel insurance. Model changes concern the claim severity distribution function f (normal, gamma or inverse gaussian), the link function g and the choice of explanatory variables x. As demonstrated in the pp-plots from Figures 3 to 5, logged claims match more closely the gamma distribution, although transforming the response variable affects the predictive performance of a model which seeks to predict the original claim variable.

In this section, the goal is to select a model with sufficient simplicity to ensure reasonable interpretation of the coefficients of regression for future application in actuarial rating methods. Furthermore, we prefer models that are statistically sound and which generalize well to new unseen data, using the test dataset. Here, we describe the candidate models which were compared using the statistical measures described in previous sections, with actuarial considerations in mind. Section 7.1 first introduces how a GLM with no explanatory variable behaves, and is used as a baseline model. This model will output a predicted value equal to the average claim for all claims:

$$\hat{y}_i = \bar{y} \ .$$

Therefore, we expect that any other model variations including additional explanatory variables or response variable transformation should improve the fitting measures and the predictive performance over the baseline model. Section 7.2 presents a simplistic GLM with log link function g, gamma distribution f and its only two significant explanatory variables. Given the few number of significant variables of this model, we present in Section 7.3 a GLM with logged claims as the response variable with 9 significant variables but with poor predictive performance of the original claim variable. To address this poor predictive performance, which is explained in part by the distinct behaviour of claims below and over 1,000$, we construct a model which uses multiple GLMs and splits claims based on their amount. We then improve significantly the model in Section 7.5 and use the log link function g with the gamma distribution f, as in Section 7.2, but remove claims above 100,000$ because they are considered outliers.

7.1 Model 1: Average claim as prediction with gamma distribution and log link function

The baseline model we use as a reference is the mean of claims from the training dataset as the claim prediction:

$$\hat{y}_i = \frac{1}{n} \sum_{i=1}^{n} y_i$$

We use this very simplistic model to introduce the illustrations and measures used to compare models with each other. For this, we used a GLM modeling the claim as the response variable, and no explanatory variables are included to force the model to compute the average claim as the prediction. The gamma distribution f was selected given its wide use for claim severity, with the log link function g to facilitate the interpretation of the coefficients of regression and derive multiplicative factors. Here is how the model is

specified:

$$y_i \sim Gamma(\alpha = 1/\phi, \beta_i = \alpha/\mu_i)$$

$$\log(\mu_i) = \beta_0 \rightarrow \mu_i = e^{\beta_0}$$

$$\hat{y}_i = \hat{\mu}_i = e^{\hat{\beta}_0}$$

Below is the corresponding output summary from the GLM model in R:

```
Call:
glm(formula = Montant ~ 1, family = Gamma(link = "log"), data = train_sub)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.6227  -1.9565  -1.5730  -0.8572  18.6479

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.80061    0.03334     234   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Gamma family taken to be 21.06198)

    Null deviance: 69222  on 18950  degrees of freedom
Residual deviance: 69222  on 18950  degrees of freedom
AIC: 314394
```

The first line of the output specifies the model components, including the response variable to model, the explanatory variables, the distribution function, the link function and the dataset used for training. The coefficients section summarizes the value of each coefficient, which in this case is limited to the intercept β_0. The predicted value is then computed as

$$\hat{y}_i = \hat{\mu}_i = e^{\hat{\beta}_0} = e^{7.80061} = 2,442.09$$

The amount of $2,442.09$ is interpreted as the average claim.

Significance codes '.', '*', '**' and '***' are meant to illustrate easily the level of significance of a coefficient of regression in the model in increasing order, and is related to the p-value. Here, the model suggests that the intercept is highly significant, and we conclude that

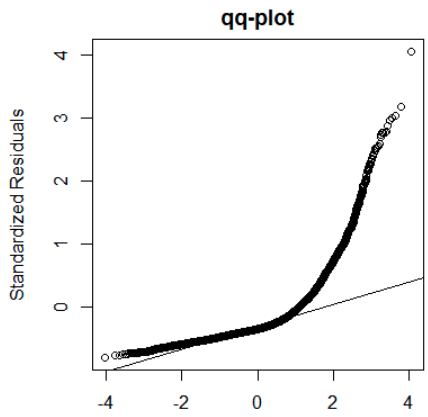
$$\beta_0 \neq 0 \ .$$

The specified Dispersion parameter is ϕ as described in previous sections, and is related to the Gamma parameters α. Therefore, for each claim, this model presumes the following:

$$y_i \sim Gamma(\hat{\alpha} = \frac{1}{\hat{\phi}}, \hat{\beta} = \frac{\hat{\alpha}}{\hat{\mu}_i}) = Gamma(0.047, 1.944e-05)$$

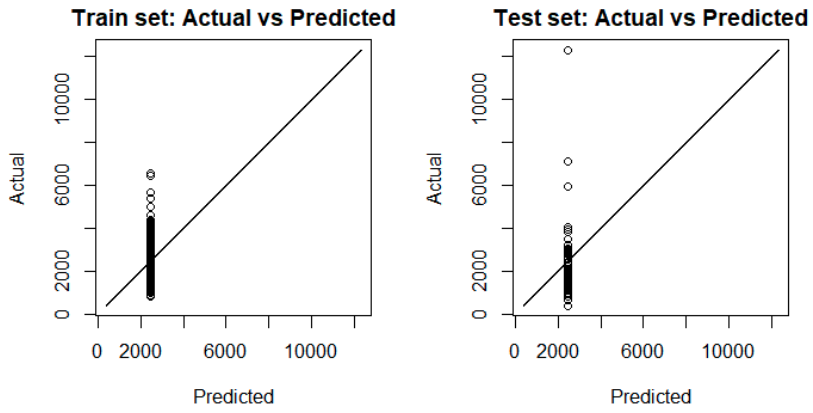
Null and Residual deviance were defined in previous sections, and are equal given that there are no explanatory variables. The AIC is $314,394$ and more complex models should target a smaller AIC value.

The qq-plot for this model is illustrated in the figure below:



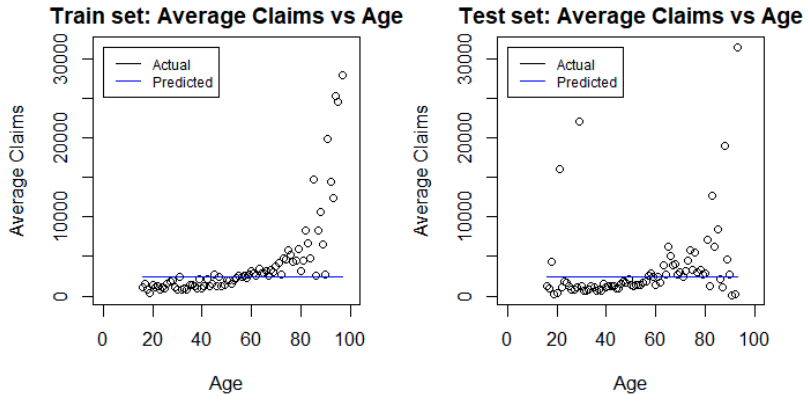
The graph above depicts if the standard normal distribution is a good fit for the the standardized deviance residuals. It visibly underlines a lack of fit in the right tail area. This suggests that larger errors are much bigger than expected, which is explained by the heavy tailed distribution of claims and the simplicity of the prediction.

Below is the graph of the actual versus predicted ratio for the training and the test set each:



Each ratio is computed for each bucket, accounting for 1% of the whole dataset ordered in increasing order of predicted values. As expected, the graph shows a vertical line given that the predicted value on the x-axis is constant and equal to the average claim of the training set. This is therefore a very bad fit and far from the target linear relation $y = x$.

The final graph displays the fit of the average predicted claims per age group to the actual average claim:



As expected again, a straight line is illustrated for the predicted values, and shows a lack of fit per age. In the next sections, we will seek models with predictions matching the depicted increasing trend of average claims per age group.

7.2 Model 2: Claim as response variable with gamma distribution and log link function

The simplest response variable is the claim, because it doesn't require any transformation. We have shown in previous sections that fitting the claims without the log transformation was a hard task, and no known exponential distribution is suited for the right heavy tail. Given that a bad fit is expected, we narrowed the choice of link function to the log to facilitate the interpretation of the coefficients. We further tested the performance of the model with the gamma, inverse gaussian and normal distributions. The only significant explanatory variables were the Age and Company factors. Below is how the R outputs for the coefficients of regression should be interpreted:

$$log(\mu_i) = \beta_0 + \beta_{age}x_{age} + \beta_{company}x_{company} \rightarrow \mu_i = e^{\beta_0} \cdot e^{\beta_{age}x_{age}} \cdot e^{\beta_{company}x_{company}}$$

$$\hat{y}_i = \hat{\mu}_i = e^{\hat{\beta}_0} \cdot e^{\hat{\beta}_{age}x_{age}} \cdot e^{\hat{\beta}_{company}x_{company}}$$

Below is a table summarizing the key measures for each distribution function f:

Model 2: Claim as response variable with log link function g
Explanatory variables: Age and company

Distribution f	AIC	RMSE Train	RMSE Test	A/P Train	A/P Test
Gamma	312,895	11,139	17,577	101%	100%
Inverse Gaussian	298,429	11,146	17,582	104%	103%
Normal	406,967	11,138	17,578	100%	100%

Model 1 from the previous section resulted in an AIC value of 314,394. The model with the normal distribution results in a much higher AIC value of 415,089, and should not be selected. The inverse gaussian appears to perform much better than the gamma with respect to the AIC value, but the predictive performance of the A/P ratio is not as convincing as for the gamma and the normal distribution. Furthermore, the inverse gaussian resulted in several convergence issues, and remains a model which is often difficult to implement.

Appendix D displays the graph illustrations related to this model with the gamma distribution. As expected, including the age factor improves significantly the fit of the predicted values per age group compared to Model 1. Furthermore, the Actual versus Predicted ratio plot depicts a trend much closer to the linear relation $y = x$. Unfortunately, Model 2 has two main pitfalls. First, the gamma is a very bad fit for the claims distribution, and a transformation of the response variable would significantly improve the fit. Secondly, there are few significant explanatory variables, including only the age and the company. Therefore, we expect a bad predictive performance per risk class for other factors such as the gender, province or salary. Model 3 in the next section suggests improvements to address these limitations.

7.3 Model 3: Logged claims as response variable with gamma distribution and identity link function

Logging the claims showed a significant improvement of fit with the exponential distributions. The GLM in this section uses the log of claims as the response variable. We express μ_i as:

$$\mu_i = E(\log(y_i)|\mathbf{x}_i) = E(z_i|\mathbf{x}_i) \ .$$

The significant explanatory variables are the age, the youth factor, the company, the employee status, the year of claim, the province, the protection type, the gender and the salary. The increased number of significant coefficients is a proof of great improvement from Model 2, which seemingly lacked complexity.

In order to recover the mean μ_{y_i} of the original variable y_i, we use the moment-generating function:

$$z_i = \log(y_i) \rightarrow y_i = e^{z_i}$$

$$\mu_{y_i} = E(y_i|\mathbf{x}_i) = E(e^{z_i}|\mathbf{x}_i) = M_{z_i|\mathbf{x}_i}(t = 1)$$

where $M_{z_i|\mathbf{x}_i}(t)$ is the moment-generating function of $z_i|\mathbf{x}_i$, and has a closed form for the normal, gamma and inverse gaussian distributions. The details to derive the mean of y_i when z_i is from a gamma distribution

are available in Appendix C. Below are the corresponding formulas for each distribution, knowing that R yields the values of $\hat{\phi}$ and $\hat{\mu}_i$ in the GLM output:

Distribution f	$M_{z_i\|\mathbf{x}_i}(t=1),\ z_i \sim f$	Parameters Estimation
Gamma	$\frac{1}{(1-\lambda_i)^{\alpha}}$	$\alpha = \frac{1}{\phi},\ \lambda_i = \frac{\mu_i}{\alpha} = \mu_i \cdot \phi$
Inverse Gaussian	$\exp\left[\frac{\lambda}{\mu_i}\left(1-\sqrt{1-\frac{2\mu_i^2}{\lambda}}\right)\right]$	$\lambda = \frac{1}{\phi}$
Normal	$\exp\left[\mu_i + \sigma^2/2\right]$	$\sigma^2 = \phi$

Using the equations above, we derive a function for the prediction $\hat{y}_i = \hat{\mu}_{y_i}$ using the estimated parameters $\hat{\phi}$ and $\hat{\mu}_i = g^{-1}(\mathbf{x}_i^{\top}\hat{\beta})$. Below is a summary of the key measures for each tested distribution function f and link function g. In these results, the AIC corresponds to the model with the response variable $z_i = \log(y_i)$, while the RMSE and the A/P ratios are related to the predictions $\hat{y}_i$ (not $\hat{z}_i$):

Model 3: Logged claims as response variable
Explanatory variables: Age, youth factor, company, employee status, year of claim, province, type of protection, gender and salary

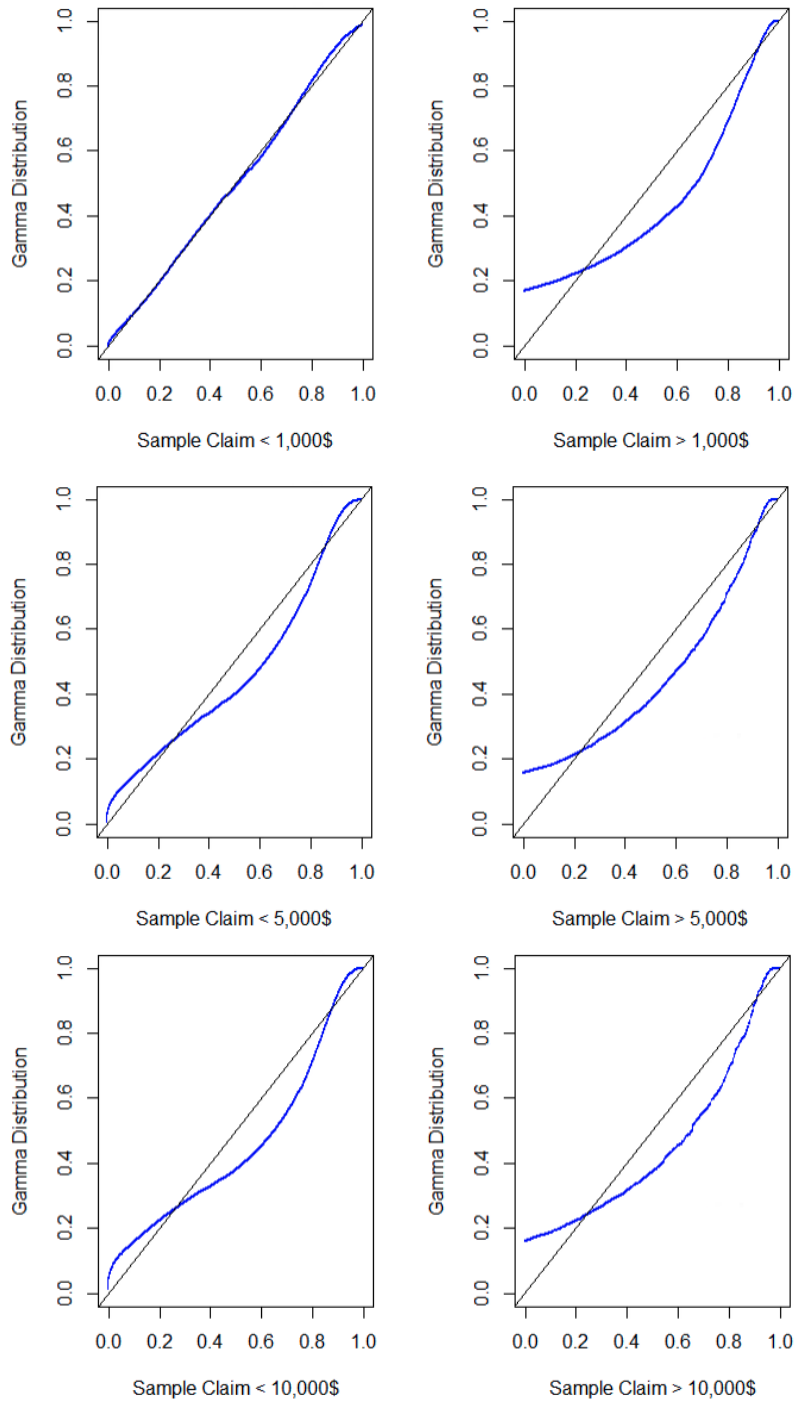
Distribution f	Link g	AIC	RMSE Train	RMSE Test	A/P Train	A/P Test
Gamma	identity	70,702	11,280	17,625	85%	85%
Gamma	log	70,702	11,402	17,633	85%	85%
Gamma	$1/\mu$	70,701	11,759	17,647	84%	85%
Inverse Gaussian	log	71,456	67,638	72,782	10%	9%
Inverse Gaussian	$1/\mu^2$	71,456	69,809	74,596	9%	9%
Normal	identity	71,967	11,215	17,620	163%	162%
Normal	log	71,966	11,214	17,620	163%	162%

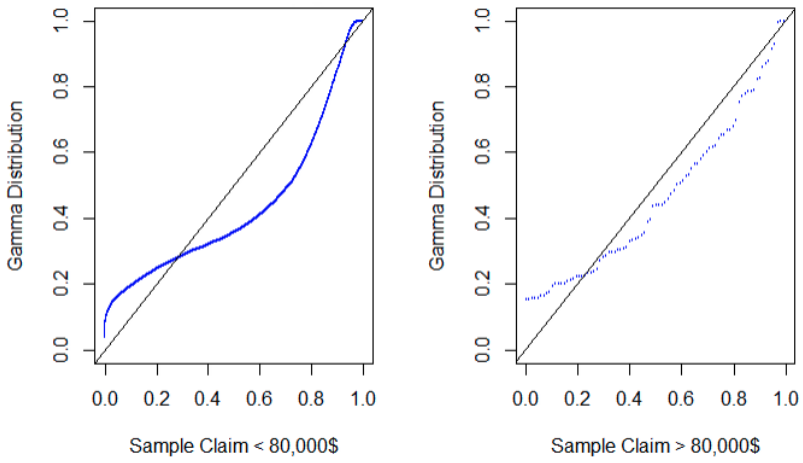
The conclusion is that all models have excellent predictions for the response variable $z_i = log(y_i)$, but they succeed poorly when it comes to the prediction of the original variable $y_i = e^{z_i}$. The gamma distribution succeeds best with an Actual versus Predicted ratio of 85% on the test set, but it is much lower than for Model 1 and 2. This current model has two advantages, one of displaying a very good fit to the logged claims and the other of having multiple significant explanatory variables. Unfortunately, the resulting coefficients are hard to interpret as multiplicative factor given the required transformation to recover the original claim variable prediction. Furthermore, results show an extremely poor prediction performance, with very high RMSE measures and Actual versus Predicted ratios far from 100%. Appendix E displays the graph illustrations of the GLM with logged claims as response variable, the gamma distribution and the identity link function.

7.4 Model 4: Splitted claims based on the amount as response variable

Results shown in the previous section suggest that a GLM can hardly fit travel insurance claims, given the heavy right tail and the lack of available explanatory variables. One reason for this is that there are probably multiple sub-populations in the database, each with distinct distributions and explanatory variables. If more detailed information about the health status of the insured or the trip were available, developing a model for each population would be a simpler task. Since these information are not available, we recommend an extremely simplistic model which defines a distinct GLM for two sets of claims which are splitted at a specific

amount. To pick this amount, an iterative process allows to define the best fit to a gamma distribution for all amounts lower or higher than the threshold amount. Below are pp-plots for different thresholds using the gamma distribution:





The best fit is observed for amounts less than 1,000$, but no fit is appropriate for larger claims. That being said, approximately 75% of claims are less than 1,000$, and therefore constitute the majority of the database. For this reason, a threshold amount of 1,000$ was selected to split the database in two sets. Two GLMs were developed using the gamma distribution and the log link function. The first is for claims less than 1,000$ and the second is for claims of 1,000$ and more.

The first model resulted in a very good fit, and the explanatory variables include the youth factor, the employee status, the year of claim, the province and the salary. An interesting result is that the age is not significant for claims less than 1,000$, which underlines that bigger claims are related to medical emergencies, which in turn are related to older ages. The GLM for claims of 1,000$ and more did not display as optimistic results, especially regarding the distribution of deviance residuals. That being said, the age, the company and the gender are the significant variables, which is a slight improvement compared to Model 2.

An additional model is required for assigning a claim to the correct group of claims. For this, a GLM with a binomial distribution and logit link is used to predict the probability that a claim is more than 1,000$ or not. The resulting explanatory variables are the age, the youth factor, the company, the employee status, the province, the protection type, the gender and the salary.

Below is how the model is defined:

$$y_{1k} \sim Gamma(\alpha_1, \beta_{1k})$$

$$y_{2l} \sim Gamma(\alpha_2, \beta_{1l})$$

$$I_i \sim Binomial(\pi_i)$$

where
y_{1k} = k-th claim $< 1,000$\$
y_{2l} = l-th claim $\geq 1,000$\$
$I_i = \mathbb{1}$(i-th claim $\geq 1,000$\$) .

The target response variable is defined as

$$y_i = y_{1i} \cdot (1 - I_i) + y_{2i} \cdot I_i$$

With the log link function and the claim amount as the response variable, we have the following linear relation:

$$\log(\mu_{1i}) = \beta_0 + \beta_{youth}x_{youth} + \beta_{status}x_{status} + \beta_{year}x_{year} + \beta_{province}x_{province} + \beta_{salary}x_{salary}$$

$$\log(\mu_{2i}) = \beta_0 + \beta_{age}x_{age} + \beta_{company}x_{company} + \beta_{province}x_{province}$$

With the logit link function and the probability of observing a claim of 1,000$ and above as the response variable, we have the following linear relation:

$$\log(\frac{\pi_i}{1-\pi_i}) = \beta_0 + \beta_{age}x_{age} + \cdots + \beta_{salary}x_{salary}$$

Below are the results of each model, with a threshold amount of 1,000$:

Model 4: Splitted claims based on the amount as response variable

1- Claim $< 1,000$\$ as response variable with log link function and gamma distribution

Explanatory variables: Youth factor, employee status, year of claim, province and salary

AIC	RMSE Train	RMSE Test	A/P Train	A/P Test
190,461	222	217	100%	99%

2- Claim $\geq 1,000$\$ as response variable with log link function and gamma distribution

Explanatory variables: Age, company and gender

AIC	RMSE Train	RMSE Test	A/P Train	A/P Test
90,059	21,331	35,386	100%	102%

3- Probability of claim $\geq 1,000$\$ as response variable with logit link function and Binomial distribution

Explanatory variables: Age, youth factor, company, status, province, protection, gender and salary

AIC	RMSE Train	RMSE Test	A/P Train	A/P Test
20,533	0.42	0.42	100%	99%

4- Mixture

AIC	RMSE Train	RMSE Test	A/P Train	A/P Test
-	11,140	17,577	100%	100%

The highlights of this model is that it is easily interpretable from an actuarial point of view, with distinct multiplicative factors for each group of claims. Furthermore, computing the prediction is fairly simple. The main pitfall is that this model is not statistically sound, and tries to replicate a mixture model using weak methods which involves splitting a database in two based on the claim amount. Furthermore, the GLM model for claims of 1,000$ and more does not result in a good fit, due to highly skewed amount on the right tail side. A more refined model would involve using a mixture model which detects all possible sub-populations with rigorous statistical methods. Nevertheless, predictions are at least as good as Model 2

and proves several advantages over the simpler models. Graph illustrations resulting from this model are summarized in Appendix F.

7.5 Model 5: Removal of claims above 100,000$

The main challenge observed in the previous models is the difficulty of finding a distribution which can fit the larger claim amount. A simple method to address this type of issue is to remove outliers, which usually improves significantly the predictive performance of a model. For this, we removed what we refer to as catastrophic claims, which are arbitrarily defined to be claims above 100,000$. Below are the statistics of claims lower than 100,000$, and higher than 100,000$:

Statistics for claims less than $100,000$\$

Min	1st Qu.	Median	Mean	3rd Qu.	Max.	Count
1.27	140.20	291.36	1985.51	879.23	98,459.99	25,206

Statistics for claims greater or equal to $100,000$\$

Min	1st Qu.	Median	Mean	3rd Qu.	Max.	Count
103,644	121,216	153,886	184,178	194,219	1,096,391	63

The GLM with a log link function and the gamma distribution excluding those catastrophic claims results in a model with the following significant explanatory variables: Age, protection type, gender, status and province. Below is a summary of the key measures of this model:

Model 5: Claims $< 100,000$\$ as response variable with log link function and gamma distribution
Explanatory Variables: Age, status, province, protection and gender

AIC	RMSE Train	RMSE Test	A/P Train	A/P Test
308,644	6,512	6,531	100%	101%

For comparison, below is the GLM model when no explanatory variable is included, which yields the average claim below 100,000$ as the prediction for all claims (same principal as for Model 1):

Model 5: Claims $< 100,000$\$ as response variable with log link function and gamma distribution
Explanatory Variables: None

AIC	RMSE Train	RMSE Test	A/P Train	A/P Test
310,088	6,579	6,595	100%	100%

Based on the AIC and the RMSE values, the model which includes all the explanatory variables displays an improvement from the model which uses no explanatory variable. It performs as well as the simpler version with respect to the Actual versus Predicted ratio, and appears as a reasonable model, although the gamma is still not a good fit. Furthermore, defining the amount at which a claim is defined as catastrophic should rely on rigorous methods and not strictly on actuarial judgement. Refer to Appendix G for the graph illustrations.

8 Conclusion

To conclude, results of the tested models have underlined the difficulty of fitting the distribution of health claims for travel insurance, and the lack of predictive performance per risk class when few explanatory variables are available. That being said, the models reveal key considerations, such as the quantitative risk related to each explanatory variable, and the need for refinement to obtain a robust model. Model 2 is the most simple one, and performs well globally, although the fit of the gamma to the claims distribution is very poor and the lack of significant explanatory variables does not reveal much information for predictive use. Model 3 shows improvements by fitting correctly the gamma distribution to the logged claims, and has multiple significant explanatory variables, but it has poor predictive performance after transformation to retrieve the original claim variable prediction. Model 4 is not statistically sound, but it is easy to understand from an actuarial point of view while including multiple significant explanatory variables. Model 5 is probably the best balance between complexity and statistical robustness, but should be improved by including an appropriate method for defining the threshold amount of catastrophic claims.

In light of these conclusions, we recommend the use of a GLM similar to Model 5 for actuarial rating purpose. Such a model presents significant improvements compared to classic group insurance rating, which is often limited to two-dimensional analysis and doesn't account for covariance of explanatory variables. Furthermore, the fitting assessment methods and predictive performance measures that were introduced in this report will be very useful for alternative GLMs for travel insurance claims, and should be used to refine actuarial calculations applied to this line of business. Another model which could surpass the performance of Model 5 is a mixture model, due to the potential presence of multiple sub-populations in the dataset. Machine learning models could also improve the predictive performance, but the lack of formal statistical equation as in the linear function in the GLM is an obstacle for actuarial analysis.

9 Appendix

Appendix A - Kernel density estimates

Kernel density estimate is a non-parametric method for displaying a smooth distribution of a continuous set of data points $X = \{x_1, x_2, \ldots, x_n\}$. A non-parametric method doesn't need to specify a closed-form expression for the data distribution (e.g. Normal, Gamma, etc...), nor to specify the underlying parameters (e.g. μ, σ for the Normal, or α, β for the Gamma). It simply depends on the set of data points. The idea is that for each value $x \in \mathbb{R}$, the contribution of point x_i to the density $\hat{f}(x)$ is relative to the distance $x - x(i)$.

The formula of the Kernel density estimate is defined as

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} K(\frac{x - x(i)}{h})$$

where K is the Kernel function and $h > 0$ is the bandwidth. The Kernel function has the following properties:

- Non-negative $K(x) > 0$
- Symmetric $\int x \cdot K(x)dx = 0$
- $\int K(x)dx = 1$

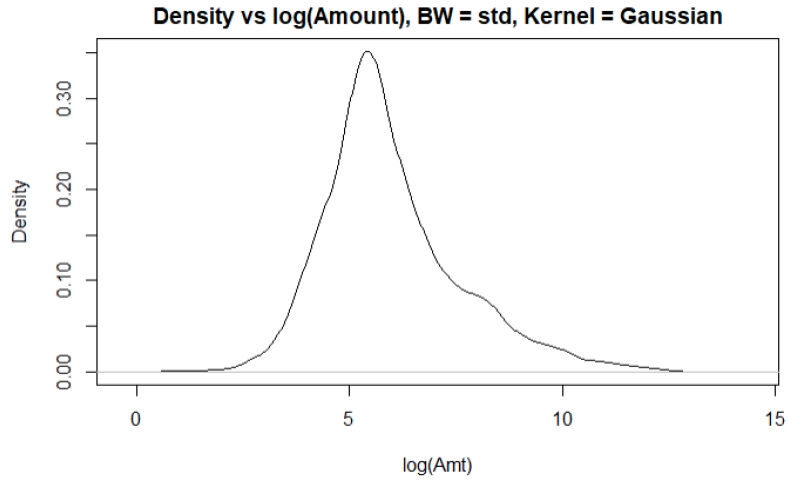
Using R, the default Kernel function used is the Gaussian:

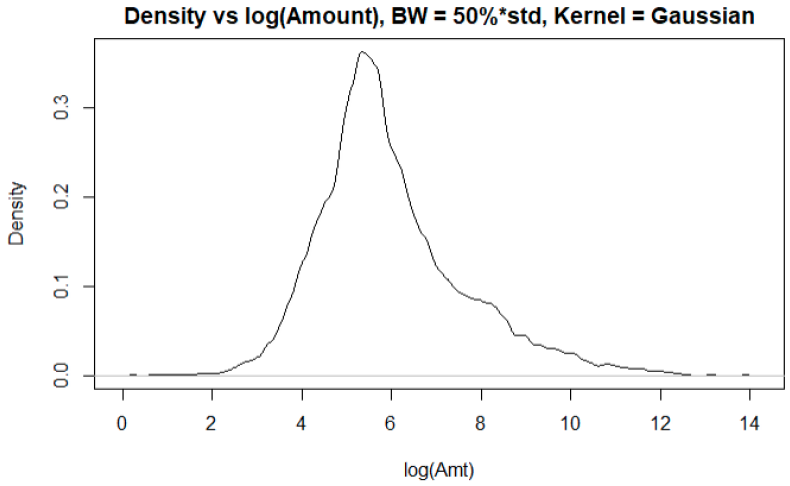
$$K_{\text{Gaussian}}(x) = \frac{1}{\sqrt{2\pi}} exp(-\frac{1}{2}x^2)$$

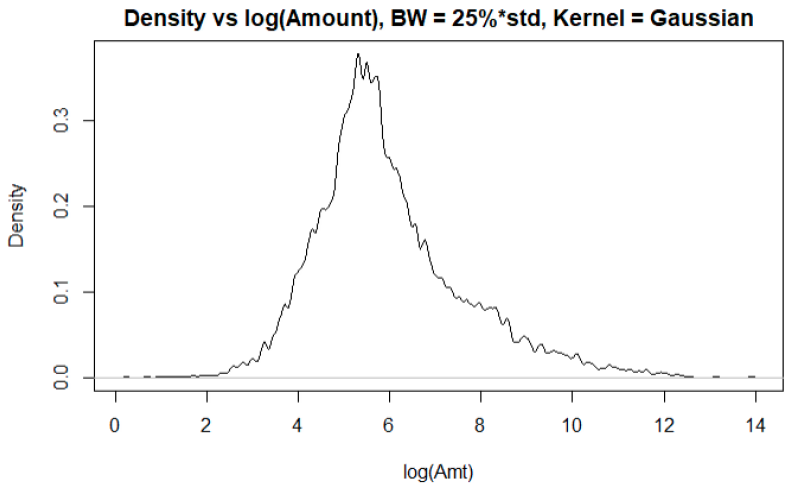
The default bandwidth is the sample standard deviation (validate). Too small bandwidth values yields little smoothing while too large values of h yields the opposite and might result in misleading conclusion.

Below are illustrations of the distribution of log-claims using Kernel smoothing with different bandwidth values and the Gaussian Kernel function (std stands for standard deviation).

Gaussian Kernel







Appendix B - Exponential family parameters estimation

The exponential family general distribution function is defined as

$$f(y;\theta,\phi) = c(y,\phi)\cdot\exp\left(\frac{y\theta - a(\theta)}{\phi}\right)$$

Closed-form formulas for $\mathbb{E}(y)$ and $Var(y)$ are derived using the two properties below:

1.
$$\int \dot{f}(y)dy = \int \frac{\partial}{\partial\theta} f(y)dy = \frac{\partial}{\partial\theta}\int f(y)dy = \frac{\partial}{\partial\theta}1 = 0$$

2.
$$\int \ddot{f}(y)dy = \int \frac{\partial^2}{\partial\theta^2} f(y)dy = \frac{\partial^2}{\partial\theta^2}\int f(y)dy = \frac{\partial^2}{\partial\theta^2}1 = 0$$

Using the first identity,

$$\dot{f}(y) = \frac{\partial}{\partial\theta} f(y) = f(y)\cdot\left(\frac{y-\dot{a}(\theta)}{\phi}\right)$$

$$\int \dot{f}(y)dy = \int f(y)\cdot\left(\frac{y-\dot{a}(\theta)}{\phi}\right)dy = \frac{\int y f(y)dy - \dot{a}(\theta)\cdot\int f(y)dy}{\phi} = \frac{E(y)-\dot{a}(\theta)}{\phi}$$

$$\frac{E(y)-\dot{a}(\theta)}{\phi} = 0$$

$$E(y) = \dot{a}(\theta)$$

Using the second identity,

$$\ddot{f}(y) = \frac{\partial^2}{\partial\theta^2} f(y) = \frac{\partial}{\partial\theta} f(y)\cdot\left(\frac{y-\dot{a}(\theta)}{\phi}\right) = \dot{f}(y)\cdot\left(\frac{y-\dot{a}(\theta)}{\phi}\right) + f(y)\cdot\frac{\partial}{\partial\theta}\left(\frac{y-\dot{a}(\theta)}{\phi}\right)$$

$$\ddot{f}(y) = f(y)\cdot\left(\frac{y-\dot{a}(\theta)}{\phi}\right)^2 - f(y)\frac{\ddot{a}(\theta)}{\phi}$$

$$\int \ddot{f}(y)dy = \int f(y)\cdot\left(\frac{y-\dot{a}(\theta)}{\phi}\right)^2 dy - \int f(y)\frac{\ddot{a}(\theta)}{\phi}dy = \frac{E[(y-\dot{a}(\theta))^2]}{\phi^2} - \frac{\ddot{a}(\theta)}{\phi}$$

$$\frac{E[(y-\dot{a}(\theta))^2]}{\phi^2} - \frac{\ddot{a}(\theta)}{\phi} = 0$$

$$E[(y-\dot{a}(\theta))^2] = \phi\cdot\ddot{a}(\theta)$$

$$Var(y) = \phi\cdot\ddot{a}(\theta)$$

Appendix C - Gamma family parameters estimation

The Gamma family density distribution function can be expressed using different parametrization. Below are the three most common forms:

- Shape and scale (α, λ):

$$f(y;\alpha,\lambda) = \frac{y^{\alpha-1}}{\lambda^{\alpha}\Gamma(\alpha)} e^{-\frac{y}{\lambda}} \tag{7}$$

$$E(y) = \alpha \cdot \lambda$$

$$Var(y) = \alpha \cdot \lambda^2$$

- Shape and rate (α, β):

$$f(y;\alpha,\beta) = \frac{\beta^{\alpha} y^{\alpha-1}}{\Gamma(\alpha)} e^{-\beta y} \tag{8}$$

$$E(y) = \alpha/\beta$$

$$Var(y) = \alpha/\beta^2$$

- GLM parametrization (α, μ) [2]:

$$f(y;\alpha,\mu) = \frac{y^{-1}}{\Gamma(\alpha)} \left(\frac{y\alpha}{\mu}\right)^{\alpha} e^{-y\alpha/\mu} \tag{9}$$

$$E(y) = \mu$$

$$Var(y) = \mu^2/\alpha$$

Here are the parameters equivalence:

$$\lambda = 1/\beta = \mu/\alpha \tag{10}$$

$$\beta = 1/\lambda = \alpha/\mu \tag{11}$$

$$\mu = \alpha \cdot \lambda = \alpha/\beta \tag{12}$$

The density function of the exponential family is defined as

$$f(y;\theta,\phi) = c(y,\phi) \cdot \exp\left(\frac{y\theta - a(\theta)}{\phi}\right) \tag{13}$$

In order to derive the canonical and dispersion parameters (θ, ϕ), the most convenient parametrization to use is the GLM (α, μ) form. Here is how equation (3) can be re-written:

$$f(y;\alpha,\mu) = \frac{y^{\alpha-1}\alpha^{\alpha}}{\Gamma(\alpha)} \cdot \exp\left(\frac{-y/\mu - \ln(\mu)}{1/\alpha}\right)$$

This way, we obtain the corresponding component of the Gamma distribution for the canonical parameter θ and the dispersion parameter ϕ:

[2] Generalized Linear Models for Insurance section 2.7

$$c(y;\phi) = \frac{y^{\alpha-1}\alpha^{\alpha}}{\Gamma(\alpha)}$$

$$\theta = -\frac{1}{\mu} \quad , \quad \phi = \frac{1}{\alpha}$$

$$a(\theta) = \ln(\mu) = \ln(-\frac{1}{\theta}) = -\ln(-\theta)$$

$$\dot{a}(\theta) = -1 \cdot \frac{1}{-\theta} \cdot (-1) = -1/\theta = \mu$$

Below is an example of R's GLM summary output:

Figure 13: GLM Summary Output

```
Call:
glm(formula = Claim ~ Variable 1 + Variable 2 + Variable 3 + Variable 4 +
              Variable 5 + Variable 6                = Gamma(link = "log"),
    data = train_sub)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.9806  -0.8320  -0.2643   0.3469   1.7793

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                    5.495e+00  2.678e-02 205.214  < 2e-16 ***
Variable 1                     3.334e-01  3.511e-02   9.497  < 2e-16 ***
Variable 2                     1.932e-01  1.751e-02  11.032  < 2e-16 ***
Variable 3                     1.570e-02  5.576e-03   2.816 0.004872 **
Variable 4                    -3.400e-02  1.457e-02  -2.333 0.019637 *
Variable 5                     4.541e-02  2.225e-02   2.041 0.041312 *
Variable 6                     7.479e-07  1.928e-07   3.879 0.000105 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Gamma family taken to be 0.670264)

    Null deviance: 11871  on 15293  degrees of freedom
Residual deviance: 11700  on 15287  degrees of freedom
AIC: 201870

Number of Fisher Scoring iterations: 6
```

The distribution function estimate of $y_i \sim Gamma(\alpha, \mu_i)$ given the output in Figure 13 are as follows

$$E(y_i|x_i) = \mu_i = g^{-1}(x_i^{\top}\beta) = g^{-1}(\eta)$$

where

$$\eta = \beta_0 + \sum_{j=1}^{p} \beta_j \cdot x_{ij}$$

$$= \text{Intercept} + \beta_{\text{Variable1}} \cdot \text{Variable1}_i + \beta_{\text{Variable2}} \cdot \text{Variable2}_i + \beta_{\text{Variable3}} \cdot \text{Variable3}_i$$

$$+\beta_{\text{Variable4}} \cdot \text{Variable4}_i + \beta_{\text{Variable5}} \cdot \text{Variable5}_i + \beta_{\text{Variable6}} \cdot \text{Variable6}_i$$

The component μ_i can also be automatically computed using the R function *predict(GLM, type = "response")*. Now using the dispersion parameter ϕ from the output, we obtain the shape parameter $\alpha = \frac{1}{\phi}$ and we have the full definition of the distribution function of $y_i \sim Gamma(\alpha, \mu_i)$.

Now let's define a variable Z such that

$$Z \sim Gamma(\alpha, \beta)$$

We further define a variable Y such that

$$Y = e^{Z}$$

The Moment-generating function of a random variable X is defined as:

$$M_X(t) := E(e^{tX}),\ t \in \mathbb{R}$$

For a Gamma distributed random variable Z, the equation with the shape and scale (α, λ) parametrization is [3]:

$$M_Z(t) = (1 - \lambda \cdot t)^{-\alpha},\ t < 1/\alpha$$

and with the GLM parametrization (α, μ),

$$M_Z(t) = (1 - \mu/\alpha \cdot t)^{-\alpha},\ t < 1/\alpha$$

Given that $E(Y) = E(e^Z) = M_Z(t = 1)$, we easily obtain the mean of the variable $Y = e^z$:

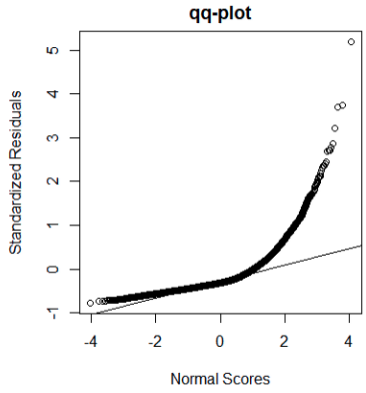
$$E(Y) = (1 - \mu/\alpha)^{-\alpha}$$

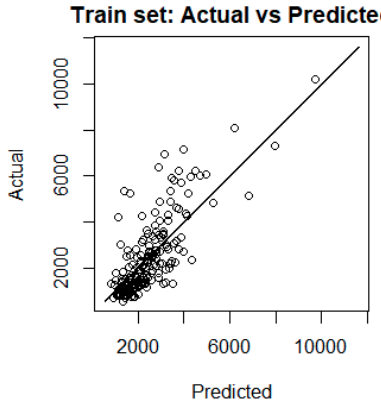
Since the GLM summary from R output gives μ and ϕ, the mean of Y can alternatively be expressed as:

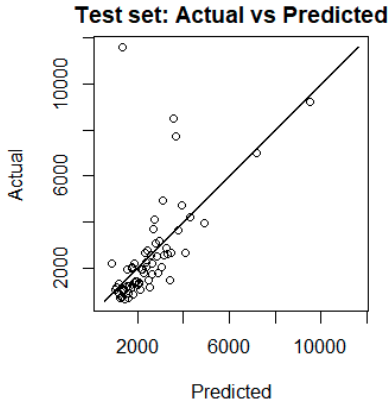
$$E(Y) = (1 - \mu \cdot \phi)^{-1/\phi} = \frac{1}{(1 - \mu \cdot \phi)^{1/\phi}}$$

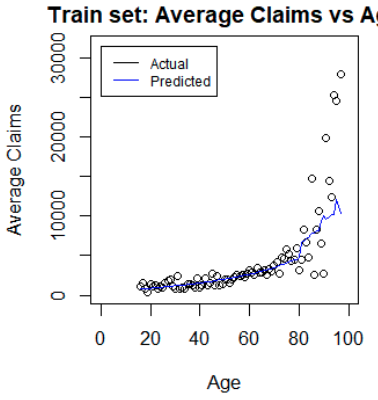
[3] https://en.wikipedia.org/wiki/Gamma_distribution

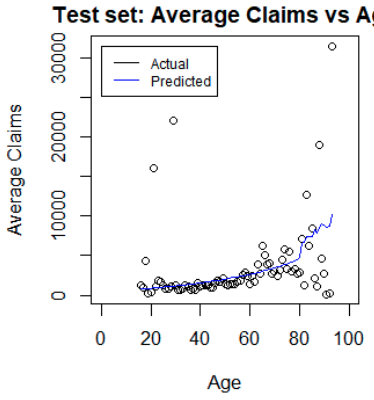
Appendix D - Model 2 Results





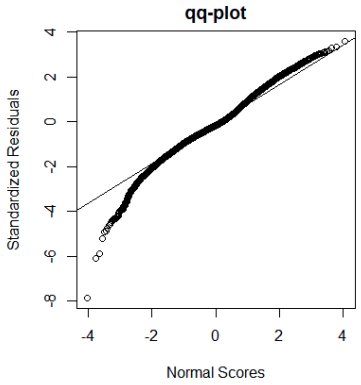


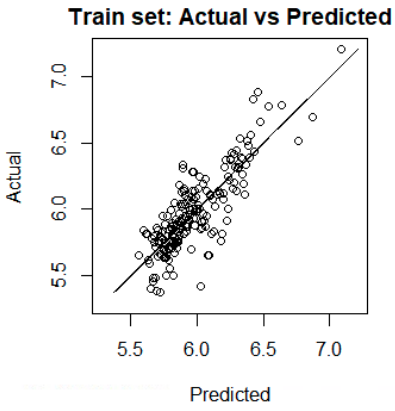


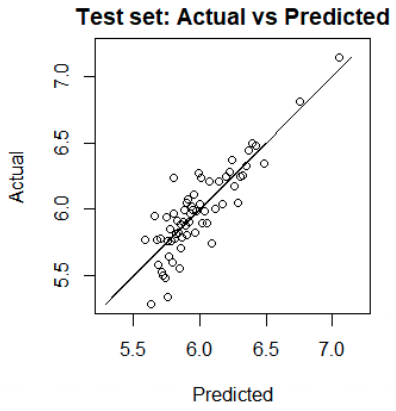


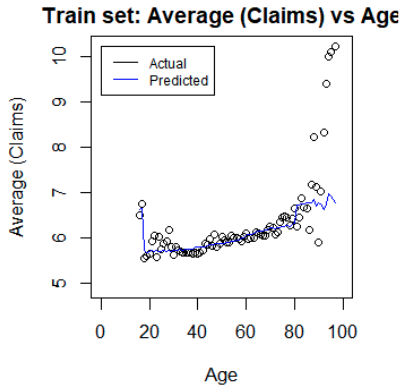
Appendix E - Model 3 Results

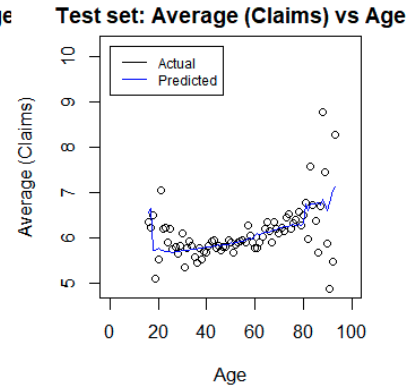
1 - Results based on $\hat{z}_i$ with $z_i = \log(y_i)$



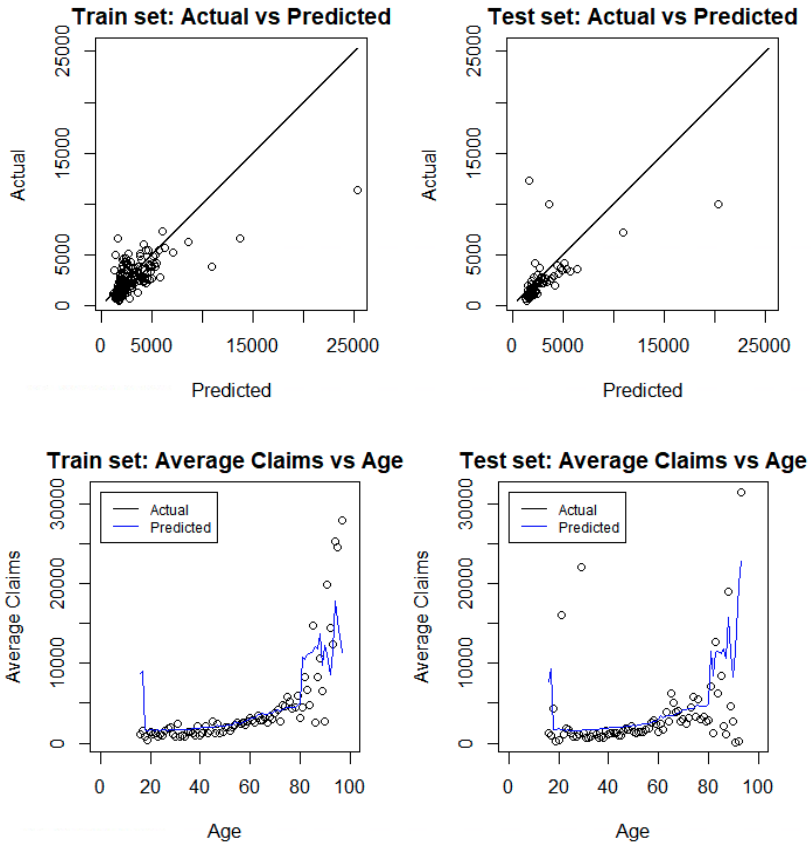






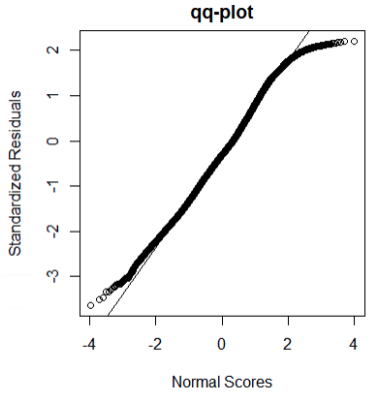


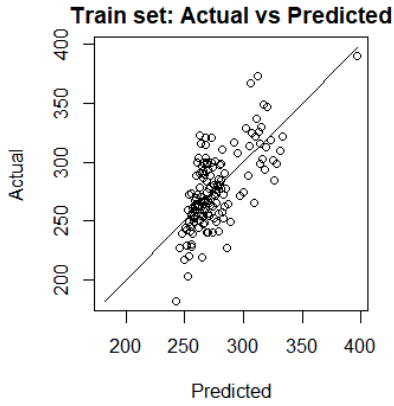
2 - Results based on $\hat{y}_i$ with $y_i = e^{z_i}$

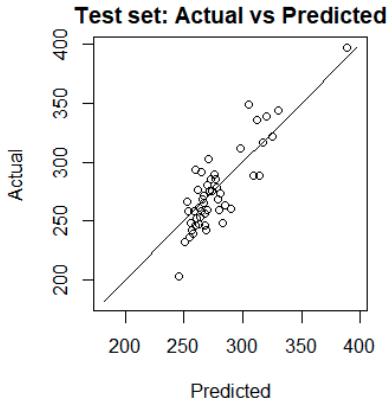


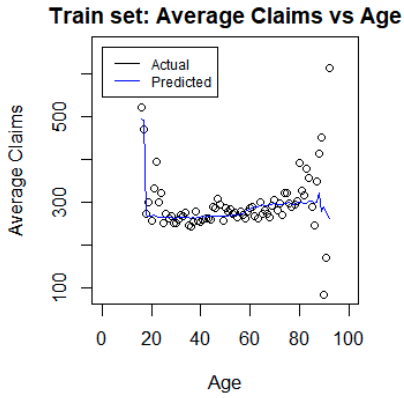
Appendix F - Model 4 Results

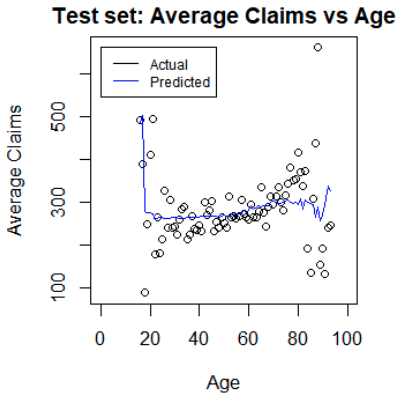
1- Results based on $\hat{y}_{1k}$ where y_{1k} = k-th claim $< 1,000$\$



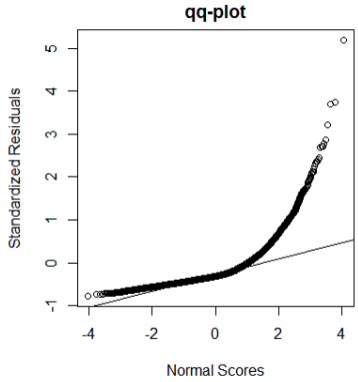


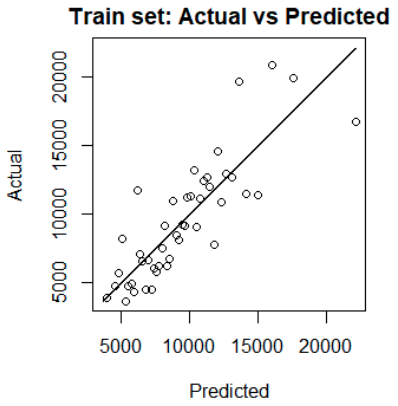


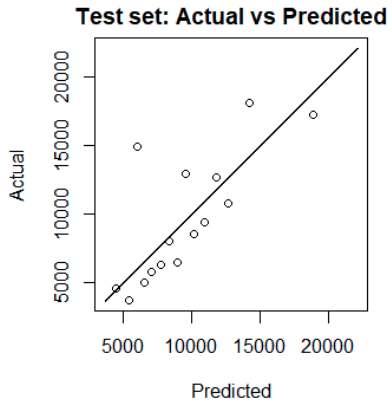


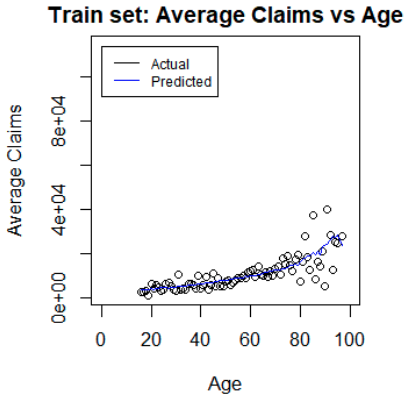


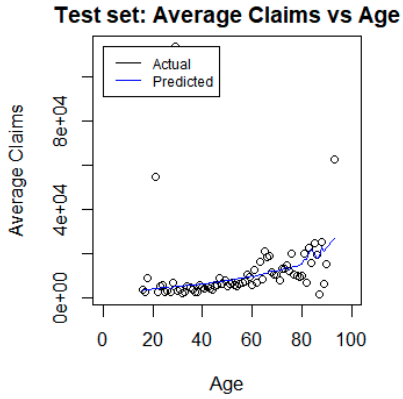
2- Results based on $\hat{y}_{2l}$ where $y_{2l} = l\text{-th claim} \geq 1,000\$$



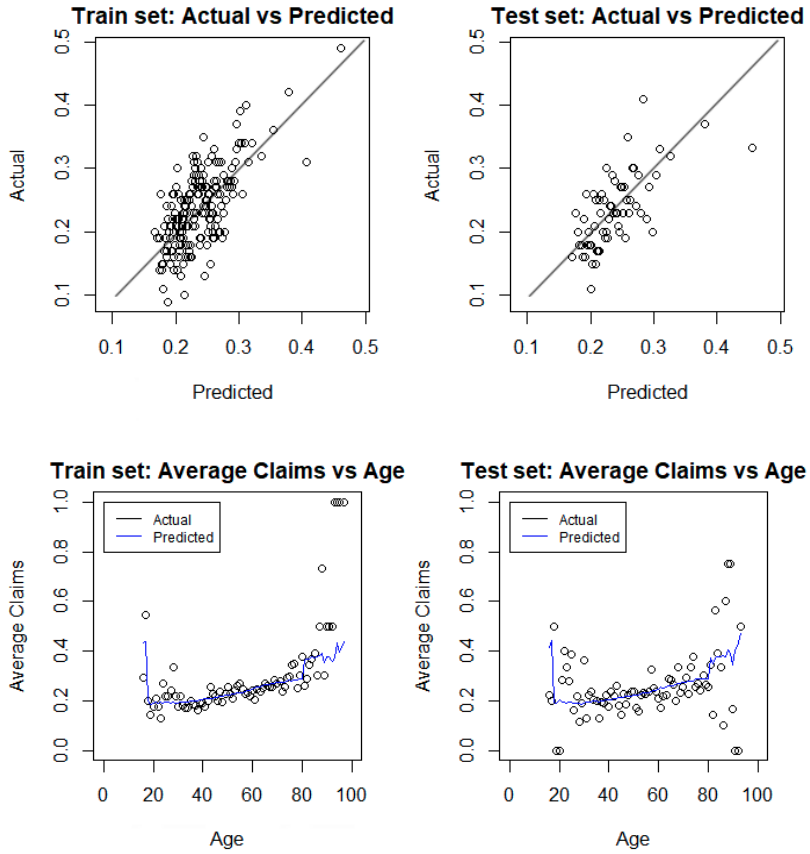




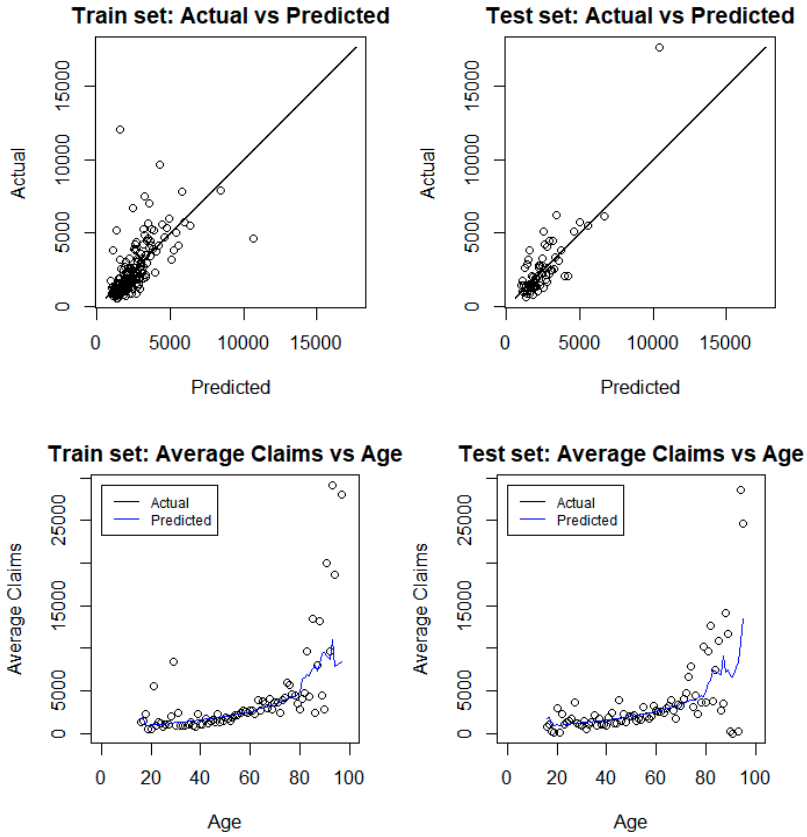




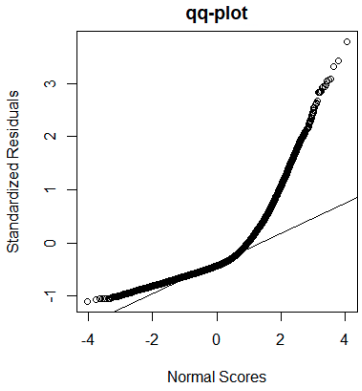
3- Results based on $\hat{I}_i$ *where* $I_i = \mathbf{1}(y_i \geq 1,000\$)$

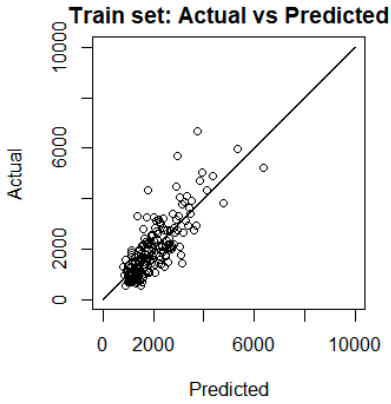


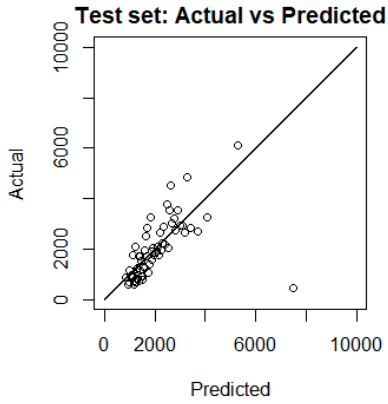
4- Results based on $\hat{y}_i$ *where* $y_i = y_{1i} \cdot (1 - I_i) + y_{2i} \cdot I_i$

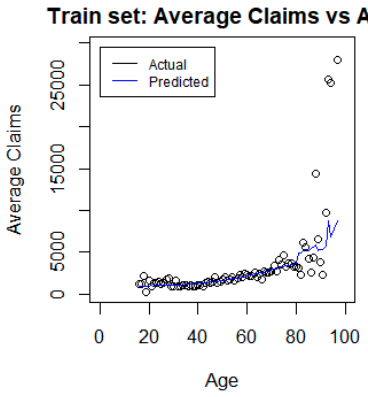


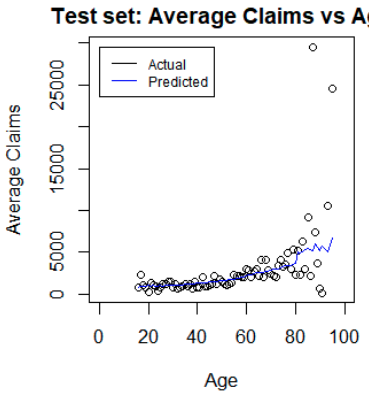
Appendix G - Model 5 Results











References

Feng, C. (2014, April). *Log-transformation and its implications for data analysis.*

Frank, S. (2016, September). *CLHIA response to CCIR's travel insurance products issues paper.*

Jong, P. D., & Heller, G. Z. (2008). *Generalized linear models for insurance data.* Cambridge University Press.

Mark Goldburd, D. T., Anand Khare, & Guller, D. (2020). *Generalized linear models for insurance rating* (Second Edition ed.). Casualty Actuarial Society.

P. McCullagh, J. N. F. (1989). *Generalized linear models* (Second Edition ed.). Chapman and Hall.

Portugués, E. G. (2020). *Notes for predictive modeling.* Retrieved from `https://bookdown.org/egarpor/PM-UC3M/`

Stone, M. (1977). An asymptotic equivalence of choice of model by cross-validation and akaike's criterion. *Journal of the Royal Statistical Society*.