

# **Best Practice Guidance Note for the Use of Generalised Linear Modelling for L&H Experience Studies**

Version 1.0

Actuarial Atelier

## **Key areas**

Generalised Linear Modelling for Actuarial Analysis for L&H business

### **Creation Date:**

26 February 2020

### **Valid from:**

26 February 2020

### **Application:**

Group-wide

### **Issued by:**

Actuarial Atelier

### **Distribution:**

Group Risk Management  
Products Underwriting L&H  
Business Management L&H

## Contents

1	Introduction .....	3
1.1	Definition .....	3
1.2	Scope.....	3
1.3	Common uses of GLM .....	3
2	Process.....	5
2.1	General analysis approach with GLMs.....	5
2.2	Key terminology.....	7
2.3	Documentation and Peer Review .....	8
3	Data.....	9
3.1	Data quality.....	9
3.1.1	Data granularity.....	9
3.1.2	Handling missing data .....	9
3.2	Data selection .....	9
3.2.1	Training dataset vs Validation dataset.....	10
3.2.2	Calibration dataset .....	10
4	Modelling .....	11
4.1	GLM considerations.....	11
4.1.1	Programming language .....	11
4.1.2	Choice of explanatory variables .....	12
4.1.3	Definition of explanatory variables .....	12
4.1.4	Probability distributions for insurance analysis.....	14
4.2	Common mistakes .....	15
5	Interpretation of results .....	18
5.1	Case Study - Continuous variable vs Categorical variable .....	18
6	Validation.....	22
6.1	Cross validation .....	22
6.2	Visual validation .....	22
6.3	Calibration .....	22
7	Reports and Communication.....	23
8	Reference .....	24
9	Version control .....	25

## 1 Introduction

This guidance note describes [Actuarial Atelier's](#)<sup>1</sup> view on the best practice in the use of a Generalised Linear Model (GLM) for Life and Health actuarial analysis. This note should be read in conjunction with the [Experience Studies Best Practice Note](#) issued by Chief Pricing Office since many aspects overlap.

This best practice guidance note aims to define the framework and high-level principles in the use of a GLM for Life and Health business to ensure consistent performance of analyses across products and geographical regions. This note is a reference point in developing a GLM and interpreting the model output for various business uses. This note will not cover every eventuality; hence we recommend discussing any areas that are not covered in this note with the Actuarial Atelier expert team and relevant subject matter experts.

GLM is widely recognised as the industry standard method for pricing of general insurance business in the European Union and many other markets. In the UK general insurance market, most insurers use a GLM regression model to determine the risk price for each peril [1]. This wide use of GLM is also observed in the US general insurance market. [2] The Life and Health insurance industry is gradually accommodating a GLM approach in actuarial analyses. Other multivariate analysis methodologies such as Cox regression are also valid as an alternative to GLM. This note aims to support the smooth transition to the use of a GLM in Life and Health analysis at Swiss Re.

Despite the many benefits of a GLM, it is simply a tool to support actuaries in making sound decisions and hence any output produced should be treated with caution and with L&H domain knowledge.

### 1.1 Definition

**A Generalised Linear Model (GLM)** is a statistical methodology that allows analysts to derive risk relationships for countless numbers of variables to predict statistical outcomes by taking interactions between variables into account and which cannot be easily done with a traditional excel-based analysis approach. A GLM is a powerful and commonly used method in predictive modelling and multivariate modelling. With GLMs, the analyst can effectively utilise information that is already available at Swiss Re, and information available externally.

For instance, a GLM can predict mortality rates as a function of any risk variable available in the data. In the GLM prediction, the model will assess the interactions between risk factors and their relationship to the predictive outcomes. Note that the GLM output should be reviewed and justified by subject matter experts. In contrast, the analyst cannot assess risks for a large number of variables and interactions effectively with a traditional excel-based approach.

A GLM is often used as a feature selection tool; it identifies key risk factors by ranking each variable's importance with statistical measures. This is an additional strength compared to excel-based analysis.

### 1.2 Scope

This guidance note focuses on the use and interpretation of a GLM in setting costing and reserving assumptions for Life and Health products. This note does not cover the GLM method for Property and Casualty business.

### 1.3 Common uses of GLM

Common uses of GLMs for Life and Health business include, but are not limited to;

---

<sup>1</sup> Actuarial Atelier is a collaboration platform tailored to the needs of actuarial teams and focusing on the application of Data Science, Automation, and Communication techniques.

- a) Derive terms of trade assumptions for mortality, morbidity, recovery and lapse behaviour etc;
- b) Analyse experience study output for costing and reserving assumption updates;
- c) Predictive underwriting;
- d) Aid underwriting decisions by identifying riskier business that is not reflected in costing; and
- e) Policyholder behaviour analysis to estimate the effect of anti-selection and selective lapses

Machine Learning (ML) is often considered in the process of a GLM since the techniques support the analyst to identify key risk factors and interactions in data without having any prior knowledge of the subject business. Due to its strong feature selection capability, ML is generally used in conjunction with a GLM to assist the initial GLM selection.

The use and interpretation of ML techniques will be included in a future update of this note. This is only applicable if ML techniques are used in the process of a GLM.

## 2 Process

### 2.1 General analysis approach with GLMs

The GLM analysis can be embedded in the traditional actuarial analysis process as an additional step. This section describes the best process for performing a GLM analysis for assumption derivations. The analyst should be clear on the purpose of the GLM analysis and plan the work by allowing for enough resources in each step. It is recommended that work in each step is documented and the rationale for any judgement is provided for peer review.

#### 1) Data preparation

Analysis data is collected and prepared for experience analysis in this step. The data volume and quality should be checked to ensure that the data is suitable for a GLM analysis. The GLM analysis is most useful when the data volume is large; the minimum data size for a GLM is often determined by pragmatic considerations. As a general rule of thumb, C Wilson et al. suggests the following guide sample sizes (i.e. number of claims, lapses etc.): 50 as very poor; 100 as poor, 200 as fair, 300 as good, 500 as very good and 1,000 as excellent [3]. This guidance needs to be considered with the intended GLM complexity since small datasets are not suitable if the analyst intends to introduce multiple risk factors and interactions. The credibility of each model point (e.g. amount of data in each risk segment) will decrease as the model becomes more complex.

#### 2) Experience analysis

The data is split by time variables (e.g. age, calendar year, policy duration), and grouped into model points. Exposures and claims are allocated to relevant time points. The experience study output should capture all risk factors at a granular level with reference tables such as current best estimate assumptions and/or industry assumptions.

The Steps 1 and 2 are the same as the traditional actuarial analysis approach. The [Experience Studies Best Practice Note](#) provides best practice in performing Step 1 and Step 2.

#### 3) Machine Learning (if available)<sup>2</sup>

In this step, the analyst performs Machine Learning (ML) modelling to obtain initial insights from the experience study output. A ML model can identify key risk factors and interactions by ranking the significance of each variable. The model can also provide an initial view on an optimal GLM based purely on statistical measures. Note that Actuarial Atelier developed<sup>3</sup> [an application](#) on Stargate that allows analysts to perform ML and GLM without any prior knowledge of coding in programming languages, such as R, Python or SAS. The application is currently named 'L&H Experience Studies 2.0'<sup>4</sup> and Actuarial Atelier plans to develop the application further to support the smooth transition to the use of data science techniques for L&H analysis.

ML is particularly useful if **a)** the underlying data is complex, **b)** the actuaries' knowledge of the data is limited (e.g. new products) and **c)** the analyst wants to monitor emerging risks that are not easily detected with the traditional analysis approach. This step is not essential for L&H business but can add more value going forward with big data.

#### 4) Traditional excel-based analysis

In this step, the analyst utilises his/her traditional actuarial analysis skills and domain knowledge to become confident with the underlying data and the ML output from Step 3. The data quality should be ensured and the ML output from Step 3 should be validated with traditional excel-based analysis (i.e. pivot table analysis and/or table analysis). The excel-based analysis is performed to identify any obvious risk factors and/or changes in the experience by utilising traditional actuarial analysis skills. If ML was not performed at Step 3, the analyst can skip the ML output validation step.

In terms of the ML output validation, the analyst should validate the ML findings against their domain knowledge to assess the reasonableness of the output. For example, assume that a ML model identifies that financial status is a significant risk factor and this risk factor interacts with sum assured amounts in a mortality portfolio; the analyst can rationalise this ML finding because both variables are a proxy for social class, and mortality risk differs by social class.

<sup>2</sup> See 'PROC GLMSELECT' and/or 'PROC HPGENSELCT' in SAS. See RANDOMFOREST and/or Lasso selection methodology in R.

<sup>3</sup> Phase 1 application tool is available as of February 2020.

<sup>4</sup> As of February 2020.

In some cases, the analyst may find the ML output contradicts his/her domain knowledge. The output needs to be studied closely to validate the ML finding.

The analyst should always rationalise likely reasons for significant risk factors and interactions identified in the ML output before using these. Where interactions and/or risk factors identified can not be explained these should not be used until rationale is available.

## 5) GLM analysis

Once the analyst is confident with the input data to a GLM (i.e. experience study output), the analyst should be clear on the purpose of the analysis before building a GLM with a preferred programming language. **Section 4.1.1.** provides a list of programming languages for GLMs and sample GLM code.

Detailed modelling considerations for building a GLM are discussed in **section 4.1.**

In this step, the analyst may split the data into training and validation datasets for output validations in Step 7. This step isn't essential for routine assumption updates<sup>5</sup> and/or if there are other validation datasets (e.g. benchmarking datasets).

## 6) Choose an optimal GLM and output validation

There are many statistical measures that help users to choose an optimal GLM. The most commonly used statistical measures include:

- BIC (**B**ayesian **I**nformation **C**riterion)<sup>6</sup>
- AIC (**A**kaike **I**nformation **C**riterion)
- Chi-Squared test
- R-squared test
- Coefficient p-values

There is no single statistical measure that is significantly better than others. BIC and AIC are the most frequently used model selection measures for a GLM, but other statistical measures should also be considered in deriving an optimal model. BIC and AIC aim to achieve a good balance between goodness-of-fit and smoothness with a penalty term. As a rule of thumb, a GLM with the smallest BIC and/or AIC is the optimal model. Although BIC and AIC are the most frequently used measures, the statistical measure that is targeted for developing a model is dependent on the purpose of the GLM.

The analyst should not solely rely on the statistical measures in deciding the optimal model since the statistical measures do not consider the qualitative aspects of the model. Qualitative aspects should be considered at this stage by considering information that is not available in the data such as product specific features and market practice, any changes in underwriting and claims control processes, admin issues etc. The selected optimal model should be cross-validated with validation datasets and calibration datasets<sup>7</sup>.

## 7) Validations

The optimal model from Step 6 should be validated quantitatively and qualitatively in this step. The most effective and powerful validation is multi-dimensional visual validation. The aim of the GLM is to explain the unexplained in the actual observations, hence the predicted values from the GLM should explain actual observations in the training dataset appropriately if the observations have a credible amount of data. The multi-dimensional visual validations should be carried out if data credibility permits. The visual validation can start from a single risk factor (e.g. by age) before a drill down to multiple risk factors (e.g. by age, gender, sum assured band). The goodness-of-fit and the smoothness of the GLM output should be checked against the actual observations for both counts and amounts (i.e. Actual vs Fitted by lives and by amounts).

Cross-validations should be performed by testing the GLM output derived from the training dataset on the validation dataset or any other benchmarking data. The validation approach is discussed in **Section 6.**

## 8) Express outputs in a table for practical uses

<sup>5</sup> This refers to routine assumption updates with a previously validated BAU GLM for the business concerned.

<sup>6</sup> Other variations include quasi-likelihood BIC and quasi-likelihood AIC

<sup>7</sup> A calibration dataset refers to a dataset that is not used in building the GLM but shares similar risk characteristics (e.g. inforce experience data can be a calibration data to test a GLM developed with pricing data).

The GLM output is generally expressed as a formula and so is difficult to judge the reasonableness of the output by looking at the coefficients alone.

We recommend converting the GLM formula to rating tables to understand the impact of each explanatory variable clearly. This will aid the communications of the output with non-GLM users. This is discussed in **Section 5.1**.

### 9) Iterative process until the final model is agreed

The Steps from 5 to 8 are repeated until the final GLM is agreed and discussed with subject matter experts. Key judgements should be documented in accordance with the [Experience Studies Best Practice Note](#).

## 2.2 Key terminology

A GLM function is expressed as below and this section explains each terminology relating to GLM.

$$g(\mu) = \beta_0 + \beta_1 * X_1 + \beta_2 * X_2 + \dots + \beta_n * X_n$$

Where:

$g(\mu)$  : Link function

$$\mu = E(Y)$$

$Y = \{Y_1, Y_2, \dots, Y_i\}$ , a range of predictions.

**Link function** ( $g$ ) links the linear form (i.e.  $\beta_0 + \beta_1 * X_1 + \beta_2 * X_2 + \dots + \beta_n * X_n$ ) with the mean of  $Y$ . Log link is the commonly used link function for a Life and Health GLM to predict a number of claims. Log link is used in this example since we assume that the model follows a Poisson distribution, which is an exponential family distribution. The log link function can be expressed as below.

$$\ln(\mu) = \beta_0 + \beta_1 * X_1 + \beta_2 * X_2 + \beta_3 * X_1 * X_2$$

$$\rightarrow \mu = \exp(\beta_0 + \beta_1 * X_1 + \beta_2 * X_2 + \beta_3 * X_1 * X_2)$$

$$\rightarrow \mu = \exp(\beta_0) * \exp(\beta_1 * X_1) * \exp(\beta_2 * X_2) * \exp(\beta_3 * X_1 * X_2)$$

**Response variable** ( $\mu$ ) is the variable that we try to predict. In a GLM, this can be number of claims, claim amount, number of lapses and lapsed amount etc.

**Explanatory variable** ( $X_1, X_2$  etc) are variables used to predict the response variable. These variables can be any variable in the data such as age, gender, sales channel and sum assured amount etc. The analyst should decide the definition of each explanatory variable at the beginning of the GLM. There are two definitions, which are continuous variables (i.e. number variables) and categorical variables (i.e. character variables). This is discussed in **Section 4.1.2**.

**Intercept** ( $\beta_0$ ) is a constant arbitrary term used to ensure that the overall GLM prediction fits the overall actual observation in the data once the effect of each explanatory variable is accounted for.

**Offset** indicates an effect in a model that is fixed by the analyst. The GLM output will depend on the choice of the offset and there are two main offset options.

#### a) Variable offset (Actual vs Expected analysis)

This fixes the effect of variables that are not being modelled, which is appropriate if the purpose of the analysis is determining relative risks to expected rates (e.g. current best estimate rates, industry tables).

Assume that the selected offset is expected claim amount and the response variable is actual claim amount in the training dataset, this GLM will perform an Actual vs Expected GLM analysis.

b) Volume offset (Rates analysis)

This reflects the fact that different model points have different data volumes (i.e. exposure). This approach is appropriate if the purpose of the analysis is constructing a new set of rates. The analyst may prefer this approach if there are no readily available expected rates and/or the actual experience is expected to deviate considerably from the expected rates. Assume that the selected offset is exposure amount and the response variable is actual amount, this GLM will perform an Actual vs Exposure GLM analysis. Therefore, the GLM output will be absolute rates (i.e. qx's, ix's etc.) not relative risks.

The choice of offset will depend on multiple factors such as data volume, types of business, main purpose of the GLM analysis and complexity of products etc.

**Confounding variables** do not appear directly in the GLM formula but should be considered when building a GLM. In general, variables are often correlated, and outcomes cannot be explained by a single risk factor. For example, if the analyst is investigating whether gender influences mortality experience (i.e. gender is an explanatory variable and mortality experience is the response variable). Confounding variables would be any other variables that influence mortality experience such as age, sum assured amount, marital status, diet, smoking status, alcohol consumption level and postcode and so on. Some of these variables are already well-understood mortality risk factors so the analyst already understands interactions between the well-known risk factors. However, this won't be the case for risk factors that are new and complex.

In this example, the analyst may find the mortality rate is higher for females compared to males when age is not considered in the GLM. Females generally live longer than males, hence the varying mortality experience by age should be considered in predicting mortality risk and its interaction with gender (i.e. females are older than males in the data, so higher average mortality rates if age is not controlled). In this situation, the analyst should control the effect of age by including the confounding variable to the model. Other effects and their interactions should also be tested. Understanding the confounding variables will result in more appropriate results.

## 2.3 Documentation and Peer Review

All steps taken to generate GLM outputs should be documented clearly and be subject to peer review. The documentation should capture the key judgements and evidence of validations. It is also considered best practice that all software and program code used to run a study be checked and subject to peer review.



## 3 Data

### 3.1 Data quality

The data input to a GLM is generally the output from an experience studies and hence the data is validated at the Experience Studies stage as per the [Experience Studies Best Practice Note](#). When building a GLM with data other than experience studies output, the analyst should perform data cleansing to perform an independent experience study as per the [Experience Studies Best Practice Note](#).

Most GLM analyses for L&H business require exposure volume (e.g. exposure lives, exposure amount) and expected volume (e.g. expected number of claims, expected amount of claim), which are attached to the data at the experience studies stage. Although data cleansing and quality assurance are performed at the experience studies stage, additional data quality assurance should be performed before building a GLM.

The quality of the GLM output will only be as good as the quality of the input data and so care should be taken in preparing and validating the input data.

#### 3.1.1 Data granularity

Some experience study output does not capture all possible risk factors and/or may be grouped with limited granularity.

To maximise the benefit of a GLM, all information should be captured at as granular a level as possible even if such information is not considered in the current assumptions. More detailed data can provide better insights in a GLM.

#### 3.1.2 Handling missing data

Missing data can be a significant problem in a GLM and will necessitate judgements in how to overcome this problem. The approach taken to cater for any such missing information can cause subsequent results to materially vary. Below are some example methods for dealing with missing information. Note that other sensible approximations can be used but the analyst must be wary that the approach adopted does not create bias in the result. It's important to consider whether the missingness is truly random or biased.

##### a) Exclude missing data from the overall analysis

It will be reasonable to exclude missing information if the impact of missing information is negligible. In other words, the missing information does not distort the overall result and the proportion of missing data is materially small [4].

If the proportion of missing information is large, excluding such data will decrease the credibility of the analysis output and may lead to bias in the results, especially if the missing information contained key characteristics.

##### b) Multiple imputations for missing value

This approach is reasonable if the missing proportion is not significant (as a rule of thumb below 40% [4]). Multiple imputations represent multiple sets of plausible values for missing data. This approach uses information from available variables to impute the likely values for the missing information. For example, assume that 10% of BMI<sup>8</sup> information is missing in a health dataset but the data quality of the other useful information, such as age, gender, health status, blood pressure etc., is good. This additional information can be used to estimate the likely values for the missing information by detecting any patterns among the known data points.

### 3.2 Data selection

Data selection will depend on the purpose of the analysis. For rates construction, the analysis period should be long enough to capture a large volume of credible data but short (recent) enough to be relevant to future

---

<sup>8</sup> Body Mass Index

experience. A good rule of thumb is 5-6 years. For a longitudinal study, the analysis period can be much longer. Most L&H experience studies are cross-sectional studies; hence the analyst focuses on the risks that were exposed during a defined time frame (e.g. analysis period x to y) rather than following the same individuals (e.g. policyholders) over a period of time. In contrast, a longitudinal study follows the same individuals over time with continuous or repeated monitoring of risk factors<sup>9</sup>, hence the analysis period can be longer for a longitudinal study compared to a cross-sectional study.

### 3.2.1 Training dataset vs Validation dataset

A **Training dataset** is the general term for samples used to create a model, while a **Validation dataset** is the samples used to validate the model created from the training dataset. As a rule of thumb, 70% of the data should be randomly assigned to the training data and the remaining 30% can be assigned to the validation data. The proportions may change, if required, but the training dataset should be larger than the validation dataset.

### 3.2.2 Calibration dataset

A calibration dataset refers to a dataset that is not used in building the GLM but shares similar risk characteristics. This dataset can be more powerful than the validation dataset to test the predictive power of the model, since this is very close to a real-world test. An example of a calibration dataset is an inforce experience dataset from the same line of business that was not used in building the GLM being tested.

---

<sup>9</sup> A specific example is 'a long-term survival analysis for those diagnosed with breast cancer in the period between 2006 and 2010'. A longitudinal study is more commonly used in a clinical research. Other examples include national studies like a census longitudinal study [15] and a birth cohort study.

## 4 Modelling

This section provides a detailed approach to building a GLM.

### 4.1 GLM considerations

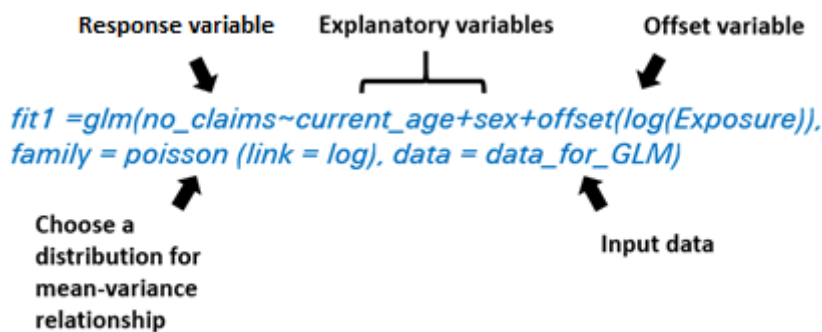
#### 4.1.1 Programming language

There are several programming languages that offer built-in GLM functionality. We provide examples of the functions available in SAS and R below.

a) [R](#)

R is a programming language and software environment for statistical computing and graphics supported by the R Foundation for Statistical Computing. The R language is widely used among statisticians and data miners for developing statistical software and data analysis. Polls, data mining surveys, and studies of scholarly literature databases show substantial increases in popularity; as of February 2020, R ranks 15th in the TIOBE index, a measure of popularity of programming languages [5] [6]. A series of R packages have been developed for advanced analytics, multivariate analysis and data science techniques (e.g. Caret, Boruta, Mlr etc.) [7]

**Glm()** is the function that tells R to run a generalised linear model. Inside the parentheses, users give R important information about the model. For example;



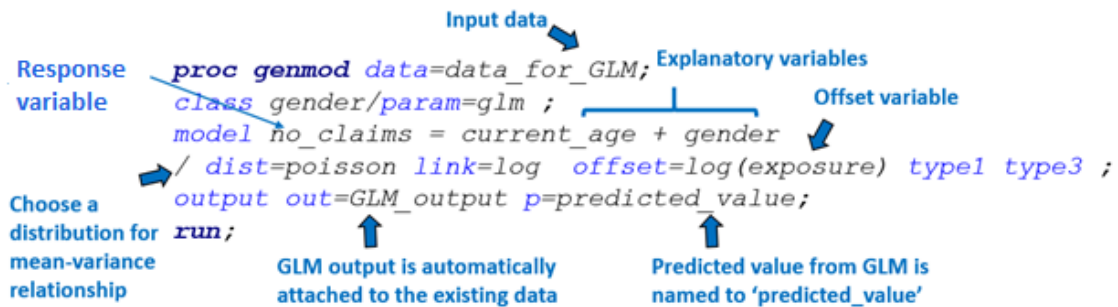
This example assumes that the response variable (no. claims) follows a Poisson distribution and the result varies by age and sex. The analyst can introduce additional risk factors and other interactions. For example, if we introduce a new risk factor 'sum assured amount' and a new interaction 'age\*gender' to the model then the formula can be written as below.

```
Fit1=glm(no_claims~current_age+sex+sum_assure_amount+current_age*sex+offset(log(Exposure))
, family=poisson(link=log),data=data_for_glm)
```

b) [SAS](#) (Statistical Analysis Software)

SAS has been developed for advanced analytics, multivariate analysis, business intelligence, criminal investigation, data management and predictive analytics. SAS provides a graphical point-and-click user interface for non-technical users and more advanced options through the SAS language. As of February 2020, SAS ranks 20th in the TIOBE index, a measure of popularity of programming languages [6].

**Proc Genmod** () is a function that tells SAS to run a generalised linear model for Life and Health data. For example,



```
proc genmod data=data_for_GLM;
class gender/param=glm ;
model no_claims = current_age + gender
/dist=poisson link=log offset=log(exposure) type1 type3 ;
output out=GLM_output p=predicted_value;
run;
```

Annotations in the diagram:

- Input data**: points to `data=data_for_GLM`
- Response variable**: points to `no_claims`
- Explanatory variables**: points to `current_age + gender`
- Offset variable**: points to `offset=log(exposure)`
- Choose a distribution for mean-variance relationship**: points to `dist=poisson`
- GLM output is automatically attached to the existing data**: points to `out=GLM_output`
- Predicted value from GLM is named to 'predicted\_value'**: points to `p=predicted_value`

Other programming languages such as Python and C++ also have built-in GLM functions.

#### 4.1.2 Choice of explanatory variables

All variables in the data are candidates for explanatory variables. The most significant risk factors are generally already known to actuaries and the choice of risk factors is not vast in Life and Health products (at least in the current underwriting setting). Having said that, the analyst should consider other potential risk factors and interactions even if they are not considered in the current assumptions.

The selection process for explanatory variables may differ by product type.

**For well-understood and well-established products**, a sensible starting point when selecting explanatory variables is to select risk factors that are currently being used. Other risk factors and interactions can then be added to the GLM to test for their significance. Note that Terms of Trade assumption can be used as an offset variable to understand deviations from the current assumptions and to identify other significant explanatory variables and interactions in the data.

**For new products**, the starting point when selecting explanatory variables is to select risk factors from similar products and/or key risk factors suggested by subject matter experts (e.g. underwriting, client markets, medical officers and Research and Development function etc.).

The analyst may find an effective new risk factor that can replace an existing risk factor. In this situation, it's recommended to investigate any potential issues such as system limitations (e.g. CPT, ICE) and commerciality (e.g. market practice) to avoid operational barriers and uncommercial decisions at a later stage. This is discussed in **section 4.2**.

#### 4.1.3 Definition of explanatory variables

The definition of each explanatory variable should be decided at the beginning of the GLM construction. Note that the analyst may change the definition at a later stage to optimise the model. There are two definitions for explanatory variables, **continuous variables** and **categorical variables**.

**Continuous variables** are number variables such as age, sum assured amount and policy duration etc.

**Categorical variables** are character variables such as gender, location and sale channel etc.

A continuous variable can be converted to a categorical variable (and vice versa) to optimise the GLM. In addition, the analyst can create two explanatory variables that have different definitions with the same source information. For example, assume that the analyst defined 'policy duration' as a continuous variable and found that the actual observation cannot be explained appropriately with the continuous variable. The analyst may create a new

variable, 'policy duration band', to treat the variable as a categorical variable. In this case 'policy duration' has two definitions with the same source information. See the case study example below.

Let's assume that the analyst found that 'policy duration' is a significant risk factor for a GLM. The initial experience investigation shows that experience by policy duration has a smooth continuous shape except for policy durations 1 and 2. The experience from the earlier duration is much heavier than longer policy durations, and the data at duration 1 and 2 is credible and they understand the likely cause of the heavy experience. This high experience also cannot be explained by other risk factors such as age, gender and sum assured etc. In other words, the heavy experience is credible and not an effect of confounding variables.

Table 1 shows the example data used to explain a practical use of a mixed variable definition. There is a continuous variable, 'dur', in the data, which captures policy duration information in a number format. Since this continuous variable does not explain the heavy experience from the durations 1 and 2 appropriately, a new categorical variable 'dur\_grouped' is created by the analyst to ensure that the GLM explains the actual observations appropriately. The number of claims information is provided in this table to demonstrate the credibility of the heavy experience from durations 1 and 2.

Policy duration in data, 'dur'	1	2	3	4	5	6
Categorical variable, 'dur_grouped'	Duration 1-2	Duration 1-2	Duration 3+	Duration 3+	Duration 3+	Duration 3+
Number of claims	1,100	1,000	600	500	200	300

**Table 1:** Example data to illustrate the mixed use of continuous variable and categorical variable

Policy duration in the data is already defined as a continuous variable. The analyst can create a new variable based on the same source information (i.e. policy duration), in this case it is named 'dur\_grouped'. Table 2 shows an illustration of the GLM output for the continuous variable and the categorical variable.

Variable definition	Variable	Coefficients	Actual vs Expected ratio to offset
Continuous variable	Dur	-0.1	0.9048 (=EXP(-0.1))
Categorical variable	Duration 1-2	0.58779	180% (=EXP(0.58779))
	Duration 3+	0.18232	120% (=EXP(0.18232))

**Table 2:** Illustration of GLM coefficients from the example in Table 1

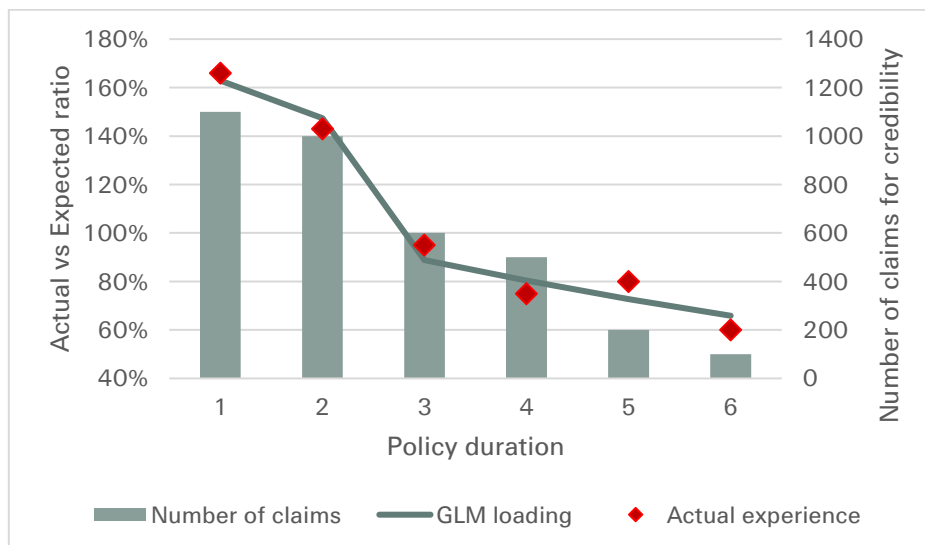
The coefficients are converted to actual loadings. In this example, it's assumed that Actual vs Expected regression is performed and thus the output is expressed as Actual vs Expected ratios.

Policy duration	1	2	3	4	5	6
GLM output for 'dur_grouped'	180%	180%	120%	120%	120%	120%
GLM output for 'dur'	90% (=EXP(-0.1*1))	82% (=EXP(-0.1*2))	74% (=EXP(-0.1*3))	67% (=EXP(-0.1*4))	61% (=EXP(-0.1*5))	55% (=EXP(-0.1*6))
GLM loading	163% (=180%*90%)	147% (=180%*82%)	89% (=120%*74%)	80% (=120%*67%)	73% (=120%*61%)	66% (=120%*55%)
Actual experience <sup>10</sup>	154%	139%	95%	75%	74%	60%

**Table 3:** Actual vs Expected ratios converted from the coefficient in table 2

Figure 1 shows that the two explanatory variables, based on the same source variable, achieve both goodness-of-fit and smoothness.

<sup>10</sup> These figures are created randomly for illustration purposes.

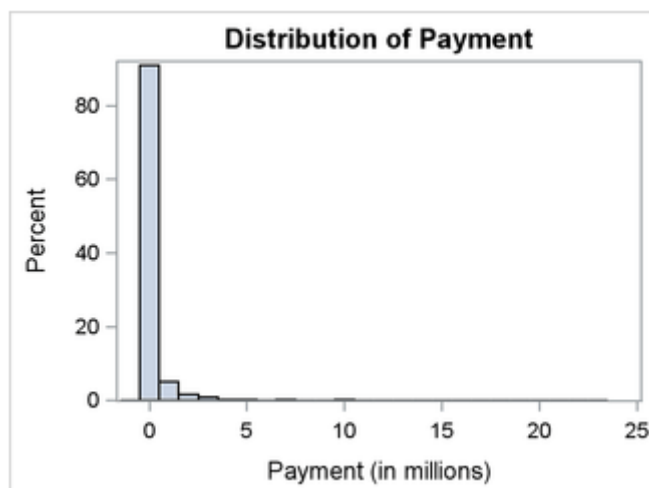


**Figure 1** : Visual validation for the mixed explanatory variable definitions

#### 4.1.4 Probability distributions for insurance analysis

There are a wide range of probability distributions for a GLM, but the most commonly used are a Poisson distribution, a Binomial distribution and a Negative Binomial distribution for frequency data and a Gamma distribution for severity data.

For an Actual vs Expected analysis on a lives basis, a poisson distribution can be selected. For Actual vs Expected analysis on a amounts basis, a mixed distribution such as **a)** Poisson-Gamma model for claim frequency and severity or **b)** a Tweedie distribution should be used to account for 2 different distributions in the data. For example, where there is a large proportion of zero amount entries from non-claimants and a distribution of claim amount from claimants, thus the observations required 2 different distributions.



**Figure 2** : An example of claim amount distribution in an insurance analysis [8]

The most commonly used probability distributions for insurance analysis are well summarised in Anderson et al. [9].

	Claim count (frequency)	Average claim amount (Severity)
Commonly used distribution	Poisson, Negative Binomial, Binomial	Gamma

<b>Link function name</b>	Log for Poisson and Negative Binomial Logit for Binomial	Log
<b>Link function</b>	$\ln(\mu)^{11}$ for Log link, $\ln(\frac{\mu}{1-\mu})$ for Logit link	$\ln(\mu)$
<b>Mean</b>	$E(Y)=\mu$ for Poisson $E(Y)=\mu$ for Negative Binomial $E(Y)=np$ for Binomial where n is number of trial and p is probability of a success	$E(Y) = K\theta$ , where K is a shape parameter and $\theta$ is a scale parameter
<b>Variance</b>	$Var(Y)=\mu$ for Poisson $Var(Y)=\mu/(1-P)$ for Negative Binomial, where p is probability a success or failure $Var(Y)=npq$ for Binomial, where s is q=1-p	$Var(Y) = K\theta^2$

**Table 4:** Probability distribution for insurance analysis, Anderson et al. and Wikipedia.

A Poisson distribution has equal mean and variance, and this causes overdispersion in the result. A Negative Binomial distribution can be used for overdispersed count data when the variance exceeds the mean. Since the variance can be larger for a Negative Binomial distribution compared to a Poisson distribution, the confidence interval will be larger with a Negative Binomial distribution, but the means should be very close. This is discussed in detail in the next section.

## 4.2 Common mistakes

### 1) Practicality matters

Analysts should bear in mind that a statistically optimal model is not necessarily an optimal model in practice. The following matters should be considered in the modelling process.

#### a) System capability

When new risk factors and/or new interactions are introduced in a GLM, the analyst should consider the system capability to ensure the newly introduced feature can be dealt with in the current system setting. For example, existing pricing and reserving tools may not be able to handle multiple interactions and the addition of a new risk factor. It may require a major system update, which can be both costly and time consuming.

The materiality of the new risk factors and/or new interactions to an existing model should be considered. If this causes system capability issues, the issue should be addressed with subject matter experts at the earliest stage.

The analyst may consider an option to account for the effect of the new risk factor or interaction at the costing stage rather than at the pricing stage to minimise potential system issues. The decision will depend on materiality and system flexibility of the proposed change with subject matter experts.

#### b) Availability of information for future business

In situations where certain information is only available from specific clients, and/or the information is no longer collected from new business, the analyst should reflect this fact in the modelling.

Let's assume that a GLM identifies 'family history' as a key risk factor based on experience data from key clients. However, a large proportion of clients do not provide the information. In this case, it will be challenging to introduce this as a new risk factor for pricing and reserving assumptions without any adjustment.

#### c) Market practice

<sup>11</sup>  $\mu$  represents the mean of the response variable (i.e.  $\mu = E(Y)$ )

Introducing a new pricing risk factor based on a GLM that other competitors currently do not include in their pricing assumptions may result in Swiss Re being exposed to higher anti-selection risk.

Market practice needs to be considered in updating the current pricing assumptions to account for the likely adverse effect in our experience due to new risk factors. This does not mean new risk factors cannot be included in pricing. The impact can be analysed in costing and its effect can be included as an explicit margin in pricing.

## **2) Over-reliance on statistical measures**

The main aim of GLMs is to help actuaries make sound actuarial decisions for actuarial analysis. A GLM is simply a tool that follows a set of statistical algorithms and hence shouldn't be used to make decisions for actuaries. Instead, actuaries should validate the output in conjunction with their domain knowledge covering commercial and practical considerations of the proposed model. Statistical measures often produce a complicated model form, but a simple model can achieve the similar statistical significance and meets commercial and practical needs better.

## **3) Overfitting**

In statistics, overfitting is defined as

*"the production of an analysis that corresponds too closely or exactly to a particular set of data and may therefore fail to fit additional data or predict future observations reliably" [10].*

Including many parameters will ensure that the GLM's predictions are close to what has been observed in the analysed data, but too many risk factors will result in overfitting with high uncertainty and hence the predictions are likely to be volatile with validation datasets (i.e. fail to predict future observations reliably).

As an extreme example, if the number of parameters is the same as the data size then a model perfectly predicts the training data. This model may not however predict future observations for other datasets.

Some statistical measures such as BIC and AIC attempt to resolve this overfitting problem by introducing a penalty term for the number of parameters in the model. The penalty term is larger in BIC than in AIC [11]. BIC and AIC will decrease as the number of parameters increase in a GLM and then start increasing when the penalty term becomes greater than the effect of the improvement of the GLM from adding additional risk factors.

## **4) Underfitting**

Underfitting occurs when a model does not capture the underlying structure of the data adequately. This is the opposite to overfitting and this GLM doesn't predict the observations in the training dataset or in the validation dataset.

This is less discussed compared to overfitting. Underfitting is generally easily detected via visual validations. For example, ignoring age as a risk factor for a mortality GLM will show a clear underfitting problem when the result is plotted by age.

## **5) Use of time variables to predict future**

Care should be taken when time variables are introduced in a GLM:

- a. A GLM does not provide credible predictions on long-term experience. A GLM simply follows the algorithm that was observed from the short-duration data. If long-duration data is limited, the GLM prediction will not be very meaningful since it extrapolates short-duration experience. Even if the underlying data contains good quality data in the long duration, the long-duration experience should also be assessed qualitatively by performing additional analysis (e.g. claims control process, benchmarking with other population etc.)
- b. Ideally, calendar year should be avoided as a risk factor as it often results in volatility. For future business, the calendar year input won't be the same and the model often simply extrapolate observed trend to future, hence this information cannot be used to predict future experience appropriately. If the analyst wants to reflect a specific period effect (e.g. unusually high deaths in one year), the adjustment should be made at the data preparation stage not at the modelling stage. If the calendar year effect still exists after controlling for all confounding factors, this should be dealt separately outside the model.



- c. Underwriting era can be considered in a GLM to reflect changes in the underwriting practices over time. Care should be taken to distinguish the effect of general risk factors with any underwriting era effect since it is likely to interact with the other risk factors (e.g. older lives with older underwriting era).

## **6) High degree of polynomial function**

Care should be taken if the GLM has a polynomial function with a high degree (e.g. cubic function, quartic function or higher) since a polynomial function with a high degree has multiple turning points (a polynomial function with 'n' degrees will have 'n-1' turning points).

Multiple turning points are generally not observed in L&H business rating structures. Having multiple turning points in the rating structure may also produce unreasonable outputs so the analyst should assess the reasonableness of the function across all key exposure areas. Attention should also be paid to the end behaviour of the function to avoid unreasonable values continuing indefinitely. Any such effect may need to be adjusted manually.

## **7) Overdispersion**

A Poisson distribution is generally selected to build a GLM for count data (e.g. number of claims). This distribution has equal mean and variance; however, the variability might be greater than expected from the statistical model. For example, if an expected value is 10, the poisson distribution expects that the variance of the observed data is also 10 but the actual variation in the data might be bigger than 10. This is known as overdispersion.

Concluding whether data is overdispersed is often reached by checking whether the ratio of the deviance to its degrees of freedom is greater than one. As a rule of thumb, if the relative variance is greater than two, then the data may be overdispersed and require statistical intervention [12].

Overdispersion doesn't generally have an impact on the mean but care should be taken if the model is used to calculate confidence intervals. Quasi-likelihood estimation is one way of allowing for overdispersion [13].

## **8) Multiple optimal models**

There is not a single right answer when choosing a GLM; multiple alternative models should be considered, and the analyst is expected to be flexible and open to challenge from market experts (e.g. Pricing, Reserving, Client Markets, Underwriting, Claims and Medical officers etc.). There may be multiple optimal GLMs that are fit for purpose. The best model is always a choice, not mechanical.

## 5 Interpretation of results

Most GLMs for Life and Health business have a log link function and therefore this section will focus on the interpretation of coefficients from a GLM with log link function. As discussed in **section 2.2**, the coefficients are expressed as below from a GLM with log link.

$$\ln(\mu) = \beta_0 + \beta_1 * X_1 + \beta_2 * X_2 + \beta_3 * X_1 * X_2$$

$$\rightarrow \mu = \exp(\beta_0 + \beta_1 * X_1 + \beta_2 * X_2 + \beta_3 * X_1 * X_2)$$

$$\rightarrow \mu = \exp(\beta_0) * \exp(\beta_1 * X_1) * \exp(\beta_2 * X_2) * \exp(\beta_3 * X_1 * X_2)$$

Where:

$\beta_0$ : Intercept

$\beta_1, \beta_2, \beta_3$ : Coefficients.  $\beta_1$  and  $\beta_2$  show the effect of  $X_1$  and  $X_2$ , respectively.  $\beta_3$  shows an interaction between  $X_1$  and  $X_2$ .

$X_1, X_2$ : Explanatory variables

### 5.1 Case Study - Continuous variable vs Categorical variable

This section discusses differences in the output interpretations depending on the definition of the explanatory variables, whether the variable is defined as a continuous variable or a categorical variable or has a mixed definition.

If the variable is defined as a continuous variable, the GLM output will have a small number of coefficients. If the variable is defined as a categorical variable, the model will produce coefficients for all input points. The coefficient is assigned to each input point; this means that the model output needs to be recalculated if a new input point is introduced.

For example, table 4 is an example GLM output for sum assured amounts, which is defined as a continuous variable. This model produces one intercept and two coefficients for the continuous variable. This result produces Actual vs Expected ratio to an offset variable with a log link poisson distribution.

Parameter	Coefficients
Intercept	0.3
Sum assured	0.0000025
Sum assured* Sum assured	-0.00000000001

**Table 4: GLM interpretation example – Sum Assured as a continuous variable**

The coefficients can be expressed as the formula below.

$$\mu = \exp(\beta_0 + \beta_1 * X_1 + \beta_2 * X_1 * X_1)$$

$$\text{Actual vs Expected ratio} = \exp(0.3 + 0.0000025 * X_1 + -0.00000000001 * X_1 * X_1)$$

where  $X_1$  is any sum assured amount.

The above formula then can be converted into a rating table to assess the effect of the sum assured and to explain the GLM result effectively to non-GLM users. The GLM output is converted to rating table for sample sum assured amounts in Table 5.

Sum assured amount	Actual vs Expected ratio
10,000	138% (=EXP(0.3+10,000*0.0000025 +10,000*10,000*-0.00000000001))
20,000	141%
30,000	144%
40,000	147%
50,000	149%
60,000	151%
70,000	153%
80,000	155%
90,000	156%
100,000	157%
110,000	157%
120,000	158%
130,000	158%
140,000	157%
150,000	157%

**Table 5: GLM interpretation example – Convert the continuous variable from Table 4 to Actual vs Expected ratio**

Since the variable is defined as a continuous variable, the model can generate Actual vs Expected ratios for any sum assured amounts. For example, the Actual vs Expected ratio for a random sum assured amount of 55,555 is 150.4% ((=EXP (0.3+55,555\*0.0000025+55,555\*55,555\*-0.00000000001)).

When the analyst interprets the result in Table 5, they should question the reasonableness of the Actual vs Expected ratio shape by sum assured amount. The Actual vs Expected ratio increases with sum assured until the amount reaches 120,000 then the Actual vs Expected ratio starts decreasing at sum assured amounts greater than 120,000. The peaked Actual vs Expected ratio at the sum assured amount of 120,000 is likely caused by the quadratic formula in the GLM function (e.g. sum assured\*sum assured in the GLM formula, see Table 4), which is imposed by the analyst. The analyst should validate the peaked shape with actual experience in conjunction with the credibility of the observations (e.g. number of claims).

If the same shape is observed in the actual experience, the analyst should be able to explain the likely cause of the quadratic shape. Note that the analyst should not introduce shapes that cannot be rationalised. For example, the quadratic shape by sum assured may be a result of some outliers (e.g. large claims), hence the shape is not suitable for predicting future experience. If the quadratic shape can't be explained, best practice is to cap the effect of the sum assured amount rather than reflecting the actual experience as it is (e.g. cap the effect of sum assured at 100,000). Note that this applies to both continuous variables and categorical variables.

If the variable is defined as a categorical variable, the GLM will generate coefficients for all input values. Converting a continuous variable to a categorical variable without any adjustment may cause a performance issue and may generate unreasonable GLM output. This may cause an overfitting issue since there are numerous sums assured values in the data, we recommend that the values are grouped appropriately by considering the experience variation and credibility of each categorical group, before changing the definition to a categorical variable. For example, the analyst may group the sum assured values as below. Once the values are grouped, the GLM will produce corresponding coefficient for each group.

Parameters		Coefficients
Intercept		0.35000
Sum assured band (SA band)	<25,000	0.000000
	25,000-50,000	0.035262
	50,000-75,000	0.062110
	75,000-100,000	0.088255
	100,000-125,000	0.150775
	125,000-150,000	0.075268
	150,000+	-0.401293

**Table 6: GLM interpretation example – Sum Assured as a categorical variable**

The coefficients can be expressed as the formula below.

$$\mu = \exp(\beta_0 + \beta_x)$$

, where  $\beta_x$  are one of the sum assured band.

Actual vs Expected ratio for SA band '<25,000' =  $\exp(0.35 + 0)$

Actual vs Expected ratio for SA band '25,000-50,000' =  $\exp(0.35 + 0.035262)$

.

.

Actual vs Expected ratio for SA band '150,000+' =  $\exp(0.35 - 0.401293)$

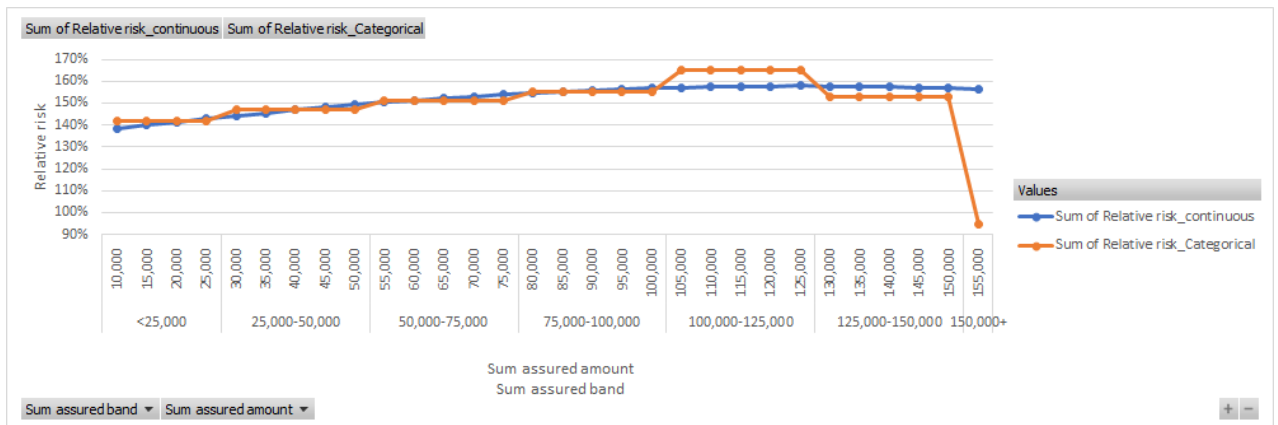
The above formula then can be converted into a rating table to assess the effect of sum assured and to explain the result effectively to non-GLM users. The GLM output is converted to rating table for sample sum assured amounts in Table 7.

Sum assured amount	Actual vs Expected ratio
<25,000	142% =( $\exp(0.35 + 0)$ )
25,000-50,000	147%
50,000-75,000	151%
75,000-100,000	155%
100,000-125,000	165%
125,000-150,000	153%
150,000+	95%

**Table 7: GLM interpretation example – Convert the continuous variable from Table 6 to Actual vs Expected ratios**

Like the GLM output with the continuous variable, the experience peaks at sum assured band 100,000-125,000. The categorical variable reflects the actual experience more closely for each group, but it introduces the risk of overfitting. For example, the Actual vs Expected ratio at the sum assured band 150,000+ drops to 95%; this sudden drop may be caused by limited data at the end group. It will also be challenging to explain the large difference in the Actual vs Expected ratio between someone with a sum assured amount of \$150,000 and another individual with a sum assured amount of \$150,001. the additional \$1 drops the risk by 60% if we use the result without any adjustment.

Compared to the result in table 5, the categorical variable produces a less smooth output. See Figure 3 for the comparison.



**Figure 3:** Model output comparison – Continuous variable vs Categorical variable

The categorical variable creates step functions in the result and overfits for some group (e.g. sum assured band 150,000+).

The analyst should take extra care when deciding the definitions of explanatory variables and the effect of the GLM coefficients should be rationalised to ensure the reasonableness of the output. The result should be cross-validated with the analyst's knowledge (i.e. the risk should not change significantly by increasing the sum assured amount by \$1.)

## 6 Validation

### 6.1 Cross validation

The underlying data can be split into subsets (e.g. by time-period, random selection) to then test the optimal GLM on each subset to ensure its reasonableness. Alternatively, the modeller can refit a GLM on each subset and compare the parameter estimates from each model to minimise model risk [14]. The dataset can be split into a training dataset and a validation dataset from the beginning for this cross-validation exercise.

### 6.2 Visual validation

Visual validation is often considered to be the most practical and effective validation method. The analyst can check the reasonableness of predicted values from the optimal GLM by plotting GLM predictions against the actual observations in the data. A good model should explain the actual observations closely (goodness-of-fit) and ensure smoothness of the predictions.

Credibility of the actual observations should be considered when validating goodness-of-fit. For example, a prediction for a business segment with a high number of claims (i.e. credible amount of data exists) should be close to the actual observation. For a business segment with scarce data (e.g. experience at extreme ages), the reasonableness of the GLM predictions should be validated by subject matter experts.

### 6.3 Calibration

Some aspects of calibration overlap with the cross-validation approach. The analyst can validate the GLM output from the training dataset against other benchmarking datasets then adjust the effect of certain coefficients to reflect any findings from the calibration result. For example, assume that the training dataset has limited information on long-duration experience so its predictive power here is limited. The output can be calibrated against benchmarking datasets that have richer data at long-duration to validate and improve the GLM output.

## 7 Reports and Communication

Any report produced from the use of a GLM should include detailed explanations concerning the methodology, including;

- Underlying data source
- Data quality assurance method
- Analysis period
- Approach in selecting training dataset and validating dataset (e.g. 70% vs 30% random selection)
- Documentation of GLM code demonstrating the choice of link function, distribution and definition of explanatory variables
- GLM output such as coefficients and statistical measures
- Details of other validation work performed (e.g. cross validations, visual validations, calibrations)
- Clear documentation on areas where judgements are exercised and the rationale

For communication of the result, the analyst should demonstrate the reasonableness of the proposed rates derived from a GLM by comparing the actual experience with the fitted result from the GLM. In demonstrating the result, the credibility of the output should be provided (e.g. number of claims in each concerned risk segment).

## 8 Reference

- [1] Deloitte, "Con-19-052 trends in general insurance pricing - Final report for the FCA," 17 May 2019. [Online]. Available: <https://www.fca.org.uk/publication/market-studies/ms18-1-2-annex-6.pdf>.
- [2] M. Goldburd, "Generalized Linear Models for Insurance Rating," Casualty Actuarial Society, [Online]. Available: <https://www.casact.org/pubs/monographs/papers/O5-Goldburd-Khare-Tevet.pdf>. [Accessed 31 January 2020].
- [3] C. Wilson, "Understanding Power and Rules of Thumb for Determining Sample sizes," [Online]. Available: <http://www.tqmp.org/RegularArticles/vol03-2/p043/p043.pdf>. [Accessed 3 February 2020].
- [4] J. C. Jakobsen, "When and how should multiple imputation be used for handling missing data in randomised clinical trials - a practical guide with flowcharts," BMC Medical Research Methodology, 6 December 2017. [Online]. Available: <https://bmcmmedresmethodol.biomedcentral.com/articles/10.1186/s12874-017-0442-1>.
- [5] Wikipedia, "R (programming language)," [Online]. Available: [https://en.wikipedia.org/wiki/R\\_\(programming\\_language\)](https://en.wikipedia.org/wiki/R_(programming_language)). [Accessed 7 February 2020].
- [6] TIOPE, "TIOBE Index for February 2020," [Online]. Available: <https://www.tiobe.com/tiobe-index/>. [Accessed 7 February 2020].
- [7] T. d. science, "Top R libraries for Data Science," [Online]. Available: <https://towardsdatascience.com/top-r-libraries-for-data-science-9b24f658e243>. [Accessed 7 February 2020].
- [8] SAS, "Fitting Tweedie's Compound Poisson-Gamma Mixture Model by Using PROC HPGENSELECT," [Online]. Available: <https://support.sas.com/rnd/app/stat/examples/tweedie/tweedie.htm>. [Accessed 3 February 2020].
- [9] D. Anderson, "A Practitioner's Guide to Generalized Linear Models," [Online]. Available: <https://www.casact.org/pubs/dpp/dpp04/04dpp1.pdf>. [Accessed 3 February 2020].
- [10] OxfordDictionaries.com, "Definition of 'Overfitting'," [Online]. Available: <https://www.lexico.com/definition/overfitting>. [Accessed 24 January 2020].
- [11] Wikipedia, "Bayesian information criterion," [Online]. [Accessed 24 January 2020].
- [12] E. H. Payne, "An empirical approach to determine a threshold for assessing overdispersion in Poisson and negative binomial models for count data," NCBI, [Online]. Available: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6290908/>. [Accessed 3 February 2020].
- [13] APES, "Introduction: what is overdispersion?," [Online]. Available: <http://biometry.github.io/APES//LectureNotes/2016-JAGS/Overdispersion/OverdispersionJAGS.html>. [Accessed 3 February 2020].
- [14] A. K. D. Tavet, "Introduction to Predictive Modeling Using GLMs - A Practitioner's Viewpoint," Insurance Programs and Analytics Services, [Online]. Available: [https://www.casact.org/education/annual/2013/handouts/Paper\\_2858\\_handout\\_1467\\_0.pdf](https://www.casact.org/education/annual/2013/handouts/Paper_2858_handout_1467_0.pdf). [Accessed 20 January 2020].
- [15] O. o. N. Statistics, "Longitudinal Study (LS)," [Online]. Available: <https://www.ons.gov.uk/aboutus/whatwedo/paidservices/longitudinalstudy>. [Accessed 7 February 2020].



## 9 Version control

Version	Detail	Date	Written by	Reviewed by
<b>Version 1.0</b>	1 <sup>st</sup> GLM best practice guidance note issued	26/02/2020	Younkyung Cho, Actuarial Control	Ian Lennox, L&H Product Centre Amit Depala, Actuarial Control