Use of generalised linear modelling in Indian insurance market for pricing health products

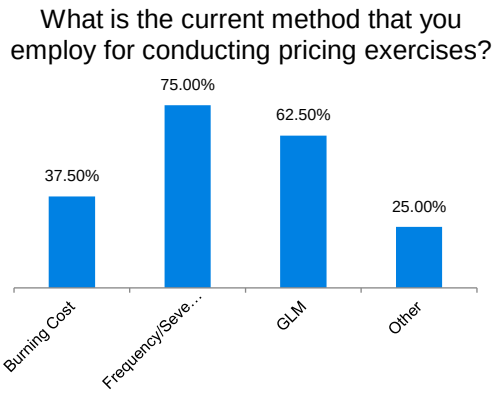
Joanne Buckle, FIA
Ankush Aggarwal, AIAI
Pravin Harodia

Introduction

Traditionally, health actuaries in the Indian insurance market used to consider generalised linear modelling[1] (GLM) to be a 'black box' for pricing of health products. However, this perception has changed significantly over time and health actuaries are now willing to employ this technique for pricing of health products.

The health actuarial team of Milliman conducted a survey to study the current and potential use of GLM in the Indian health market. All the eight respondents of the survey were interested in using GLM for future pricing purposes. The graph in Figure 1 illustrates a summary of the eight responses with respect to current use of different approaches for pricing of health products (some respondents use more than one technique). It is evident that frequency severity approach has been the most popular choice. However, GLM has also been used extensively for pricing.

FIGURE 1: SURVEY QUESTION: PRICING EXERCISE METHODOLOGY



In the survey, many insurers proposed employing additional types of validation techniques in addition to the usual validation methods currently being used. These additional types of validation techniques could be Gain curve or Gini coefficient analysis. The survey also concluded that, apart from product pricing, GLM is used by some insurers for other analysis like lapse/renewal rate modelling, large claim analysis, customer lifetime value modelling etc. Increasing numbers of insurers are interested in trying out GLM for these analyses in near future given the availability of sufficient amount of data and expertise.

[1] GLM is a technique that uses a scientific basis for identifying statistically significant relationships between a set of risk factors (explanatory variables) and the outcome variable of interest (also referred to as a response variable). In a health insurance context, response variables could include claim frequency, average cost per service, per member per month (PMPM) cost or lapse rates. GLM results can be used to predict future experience based on past claims and lapse experience of insurance policies.

The current regulations and guidance notes by the Insurance Regulatory Development Authority of India (IRDAI) do not specify the use of GLM for health insurance pricing. In addition, there are no formal best practices and guidelines defined by industry experts on conducting a GLM exercise. However, there are certain aspects of GLM that are consistently used by actuaries across the health insurance industry. An example is with respect to the selection of the distribution for frequency and severity parameters. Most of the health actuaries use Poisson distribution for fitting of frequency data and gamma distribution for fitting of severity data. Additionally, health actuaries sometimes use an interaction term of age and gender because of the different age-wise frequency and severity relativities for males and females.

In general the process of a GLM exercise can be broadly divided into four sequential steps. However, as GLM is an iterative process, the user may need to back to a previous step as part of the process

1. Developing a multivariate table for conducting GLM analysis.
2. Pre-analysis on multivariate table before conducting GLM.
3. Fitting a distribution and running GLM.
4. Reviewing the outputs of GLM and model validation.

For developing a summary table, the inputs required are the claim and member data sets of the portfolio. Necessary adjustments should be made to the data fields after checking and validating the data. On the claim side, the incurred but not reported (IBNR) factors need to be applied for completion of data. If using multiyear data, all claims need to be brought to the same level or year to be used as a factor in the GLM model to account for any trend over time. On the member side, the exposure to risk may need to be broken into per month exposure. After making all the required modifications, the claim data needs to be mapped to the member data to obtain the data set, which is often referred to as a multivariate table. This table is used as the input of GLM.

After obtaining the multivariate table, the next step is to check the overall reasonability of one-way and two-way relativities for different risk factors. Subsequently, the bucketing of the levels of the data fields should be done, for example bucketing individual ages into age bands, based on the exposure present at each level. This helps in making the data more credible at a model point level. Following the bucketing of the data, it is important to check for correlation amongst different variables to identify degrees of independence and to remove variables which are highly correlated with other included variables.

After finalising the list of variables, the distribution should be allocated based on the fit of the data (Poisson or negative binomial for frequency; gamma or inverse Gaussian for severity; Tweedie for burning cost). Furthermore, the existence of any interaction[2] of variables should be checked, and incorporated in the equation of the model.

The data should be run through the GLM process and the significance of variables in the model should be checked by doing statistical testing (tests could include chi-squared tests and comparison of Akaike information criterion and Bayesian information criterion values). Estimates of each level for a particular factor could be compared with each other using expectations and standard deviation to see the significance among different levels. This would provide more insight in defining the relevance of a factor and the type of bucketing. Residual plots can be used to check the appropriateness of the model structure like distributional assumption or incorporating mostly all the pattern. Cross-validation techniques, like 'Lift curve' or 'Gain curve,' could also be used to the test the effectiveness of the model by seeing its predictability on out-of-sample experience. It can be done by splitting the data into training and test subsets.

Finally, for pricing purposes, it is important to graduate the relativities of a risk factor to obtain a smooth progression for certain factors such as age bands and sum insured. Polynomial curve fitting or piecewise curve fitting could be used for obtaining a smooth progression of the parameter estimates.

Many software programs provide solutions for running GLM, but the most common ones used by health actuaries are SAS and R. Both of these programs have specific GLM packages ('genmod' for SAS and 'glm' for R). SAS has some advantages over R in cases of better data management and more user-friendliness. R is not able to handle big data itself as compared to SAS unless other open source programs, like SPARK, are used to boost its memory. However, in the case of R there is availability of more data visualisation and modelling capability as compared to SAS. Therefore, it is common for actuaries to first summarise data in SAS and then do GLM analysis in R.

There are many advantages of using GLM. First of all, it is able to provide relativities that can be readily used in the structure of a rate classification system. Also, in cases where there is thin data for a particular variable which is not categorical, GLM is able to predict values for that particular variable with reasonable accuracy. In addition to the best estimate, GLM can provide upper and lower estimates as part of confidence intervals to indicate the inherent uncertainty in the output. For a standard GLM process (i.e., without inclusion of any complex interactions), the computing power required to conduct GLM is not very intense.

However, there are certain disadvantages associated with using GLM; for example GLM may not always provide results that are in line with the future, especially if movements of internal and external factors are volatile. Furthermore, it is practically impossible to find all possible combinations of three or more ways of interactions in the model using trial and error method. There are techniques of using a GLM tree to identify significant interactions amongst different factors, but this process is quite time-consuming. Additionally, users of GLM make certain implicit assumptions, which may not hold true in every scenario. These assumptions could be the assignment of specific link and error functions that are required as part of the modelling process without testing the feasibility of other functions.

The table in Figure 2 summarises the advantages and disadvantages of using GLM.

FIGURE 2: ADVANTAGES AND DISADVANTAGES OF USING GLM

Particulars	GLM's status
Interpretation of results	Easy
Accuracy	Reasonably accurate
Computing power	Not very intense
Implementation within the current insurance structure	Easy to implement
Relationships between predictors & response	Assumed to be linear
Relationships between predictors	Assumed to be independent
Outliers	Very sensitive
Overfitting	Some risk
Identification of interactions	Difficult to identify automatically

Hence, the question arises as to finding a solution for addressing specific issues related to GLM. One approach would be to continue using GLM and supplementing the drawbacks of GLM by testing the assumptions used in the model and using different nonparametric approaches to identify the interaction terms. Another approach would be to discard GLM and use machine learning techniques such as neural networks, random forests and decision trees, and some ensemble methods. However, interpreting the outputs of machine learning is commonly more challenging. The use of machine learning in the Indian health insurance domain is still in its infancy, and it is quite likely that health actuaries would continue to use traditional or alternative versions of GLM for pricing purposes in the near future.

[2] Interaction is different from correlation. For more details on the difference refer to article 'Illustration of the difference between correlation and interaction amongst independent variables' by Rajagopalan Ranganathan, available at https://cnx.org/contents/1HHs9Tkt@1/Illustration-of-the-difference-between-correlation-and-interaction-amongst-independent-variables.

References

GLM II: Basic Modeling Strategy

https://www.casact.org/education/specsem/f2008/handouts/Modlin.pdf

Loss Cost Modeling vs.Frequency and Severity Modeling

https://www.casact.org/education/rpm/2012/handouts%5CSession_4738_presentation_1068_0.pdf

Introduction to Generalized Linear Models

http://statmath.wu.ac.at/courses/heather_turner/glmCourse_001.pdf

Introduction to Predictive Modeling Using GLMs
https://www.casact.org/education/annual/2013/handouts/Paper_2858_handout_1467_0.pdf

Generalized Linear Models for Insurance Rating

https://www.casact.org/pubs/monographs/papers/05-Goldburd-Khare-Tevet.pdf

Milliman

Milliman is among the world's largest providers of actuarial and related products and services. The firm has consulting practices in life insurance and financial services, property & casualty insurance, healthcare, and employee benefits. Founded in 1947, Milliman is an independent firm with offices in major cities around the globe.

milliman.com

CONTACT

Joanne Buckle
joanne.buckle@milliman.com

Ankush Aggarwal
ankush.aggarwal@milliman.com

Pravin Harodia
pravin.harodia@milliman.com