SOCIETY OF ACTUARIES®

SOA Predictive Analytics Seminar – Hong Kong

28 Aug. 2019 | Hong Kong

Session 5

GLM and its application in life insurance

Anilraj Pazhety

GLM and Applications in Life Insurance

ANILRAJ PAZHETY
Data Innovation Specialist
28th August 2019

SOCIETY OF ACTUARIES

Agenda

- Why use GLMs?
- Introduction to GLMs
- Estimating value of coefficients
- Interpreting model summary (R Code)
- GLM Applications in Life Insurance

Why use GLMs ?

$$g(\mu_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}.$$

GLMs are a powerful statistical tool to examine the underlying reality of the variables in the data in a way that can show how previously ignored variables may have an impact on the risk

Traditional statistical techniques cannot accommodate the correlations in the data between variables. This will make the one-way univariate analysis incorrect

→

GLMs capture the variations introduced due to time trends in the data thereby allowing the usage of data without restriction on the number of years of data used for analysis

→

Ability to incorporate interactions between factors make GLMs more robust and accurate

Generalized Linear Model

$$g(\mu_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}.$$

Random Component

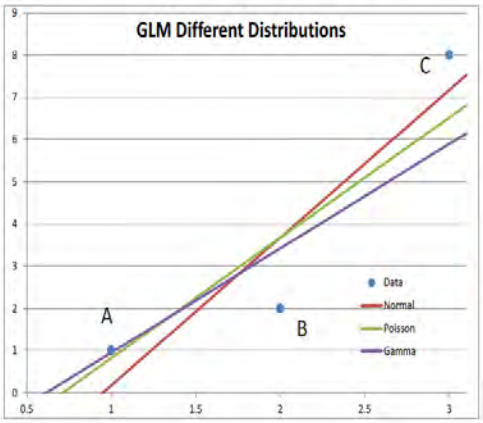
$(y_i) \sim \text{Exponential}(\mu_i)$

- μ → Mean of the distribution
- ϕ → Dispersion parameter [related to variance]

- "Exponential" is a placeholder for any member of the exponential family
- Each member is defined by parameters - μ and ϕ

	Normal	Poisson	Binomial	Gamma	Inverse Gaussian
Name	$N(\mu, \sigma^2)$	$P(\mu)$	$B(m,\pi)/m$	$G(\mu, \nu)$	$IG(\mu, \sigma^2)$
Range	$(-\infty, +\infty)$	$(0, +\infty)$	(0,1)	$(0, +\infty)$	$(0, +\infty)$
$b(\theta)$	θ^2	e^{θ}	$\ln(1+e^{\theta})$	$-\ln(-\theta)$	$-(-2\theta)^{1/2}$
$\mu(\theta)$	θ	e^{θ}	$e^{\theta}/(1+e^{\theta})$	$-1/\theta$	$(-2\theta)^{-1/2}$
$V(\mu)$	1	μ	$\mu(1-\mu)$	μ^2	μ^3

Distribution impacts the model fit



Variance of different distributions

- Gaussian, constant
- Poisson, ~ mean
- Gamma, ~ mean^2

Popular Applications of GLM Distributions

Distributions for Severity

- Gamma and Inverse Gaussian distributions have characteristics exhibited by claims severity

Distributions for Frequency

- Poisson and Negative Binomial are used for modelling expected claim count per unit of exposure or per dollar of premium

Distributions for binary target variables

- Binomial distributions are used to model the occurrence or non-occurrence of an event in a GLM model

Generalized Linear Model

$$g(\mu_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}.$$

Systematic Component

GLM models the relationship between μ_i (the model prediction) and the predictors as follows:

$$g(\mu_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \ldots\ldots\ldots\ldots + \beta p x_{ip}$$

The transformation of μ_i is represented by the function g(.) , called the link function which is a smooth & invertible function

	Identity	Log	Logit	Reciprocal
$g(\mu_i)$	x	$\ln(x)$	$\ln(\frac{x}{1-x})$	$1/x$
$g^{-1}(\eta_i)$	x	e^x	$\frac{e^x}{1+e^x}$	$1/x$

Log is unique in insurance application - all parameters are multiplicative

- $\ln \mu = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \ldots\ldots\ldots\ldots + \beta p x_{ip}$

  $= e^{\beta_0} * e^{\beta_1 x_{i1}} * e^{\beta_2 x_{i2}} * \ldots\ldots\ldots e^{\beta_p x_{ip}}$
- Consistent with most insurance practices
- Intuitively easy to understand and communicate

SOCIETY OF ACTUARIES

Statistical Methods – Coefficients Estimations

SOCIETY OF ACTUARIES

3 Key Statistics

Standard Error	▪ SE indicates how close to true value is the coefficient estimated by GLM ▪ A large standard deviation indicates that a wide range of estimates could be achieved and hence the estimate is less likely to be close the true value
P Value	▪ Estimate of the probability of a value of given magnitude arising by pure chance ▪ Lower p-values indicates that the odds of getting the estimated value by GLM is small
Confidence Interval	▪ Specifies the range of values which would be accepted at chosen p-value threshold

Offsets

- The scope of most of GLMs projects in insurance is to update some elements of the pricing basis or rating plans and not to update the entire basis at once
- In such cases, GLMs should know not to estimate coefficients for the fixed or unchanged component . Offset function takes care of this requirement
- This ensures that the coefficients of other variables are optimal in it presence

$$g(\mu_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \textit{offset}$$

Note: Offset needs to be on the scale as the other linear predictors

GLM Model Summary

Typical R code used in the process

```
Call:
glm(formula = Claim ~ Salary_Band + Occupation + Age_Last, family = "poisson",
    data = dt_df, offset = log(Expected_Deaths))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.1236  -0.0830  -0.0677  -0.0477   5.1438

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.7561993  0.1372797   5.508 3.62e-08 ***
Salary_Band2  0.0332638  0.0493582   0.674  0.50036
Salary_Band3  0.0030769  0.0530109   0.058  0.95372
Salary_Band4 -0.0004325  0.0509793  -0.008  0.99323
Salary_Band5 -0.0676021  0.0760530  -0.889  0.37407
Occupation2  -0.1858997  0.0475697  -3.908 9.31e-05 ***
Occupation3  -0.1752951  0.0434638  -4.033 5.50e-05 ***
Age_Last     -0.0109579  0.0034160  -3.208  0.00134 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 45452  on 1499999  degrees of freedom
Residual deviance: 45404  on 1499992  degrees of freedom
AIC: 52572

Number of Fisher Scoring iterations: 8
```

Step 1 - Fetch the model coefficients

```
> modl_ratings = modl$coefficients
> modl_ratings
  (Intercept)  Salary_Band2  Salary_Band3  Salary_Band4  Salary_Band5   Occupation2   Occupation3      Age_Last
 0.7561993318 0.0332638072 0.0030768505 -0.0004325044 -0.0676020939 -0.1858996579 -0.1752950595 -0.0109578539
```

Step 2- Get the exponential form of the coefficients

```
> modl_ratings_exp= exp(modl$coefficients)
> modl_ratings_exp
 (Intercept) Salary_Band2 Salary_Band3 Salary_Band4 Salary_Band5  Occupation2  Occupation3     Age_Last
   2.1301648    1.0338232    1.0030816    0.9995676    0.9346323    0.8303569    0.8392094    0.9891020
```

Note:- Expected_Deaths = base_rate *exposure

GLM Applications

SOCIETY OF ACTUARIES

Pricing Basis Review and Updates

Background

- Include more recent data to include higher number of deaths and few more years of exposure data
- Verify if the existing rating factors are appropriate and make adjustments if required
- Identify significant variables for pricing basis
- Derive additional insights beyond the traditional experience study
- Examine trends based on specific variables – Product type, Distribution Channel, City etc.

Pricing Basis Review and Updates

Data

- 8 M policies with 18M exposure counts
- >50 variables available in data set
- ~40K Death claims

Data Manipulation

- Missing Values
- Duplicates
- Decoding
- Grouping
- Binning

Applicable to all data projects

Data Transformations

- Log Link
- Poisson distribution

Pricing Basis Review and Updates

GLMs made an impact on the pricing approach

Model Validation

- **AIC** used for comparing different versions of the models
- **Lift Plots** used to measure the validity of the model based on A/E

Key Findings

- Differentiation by distribution channel and annual incomes
- Identified high risk cities
- Identified and quantified interaction between risk factors

Business Impact

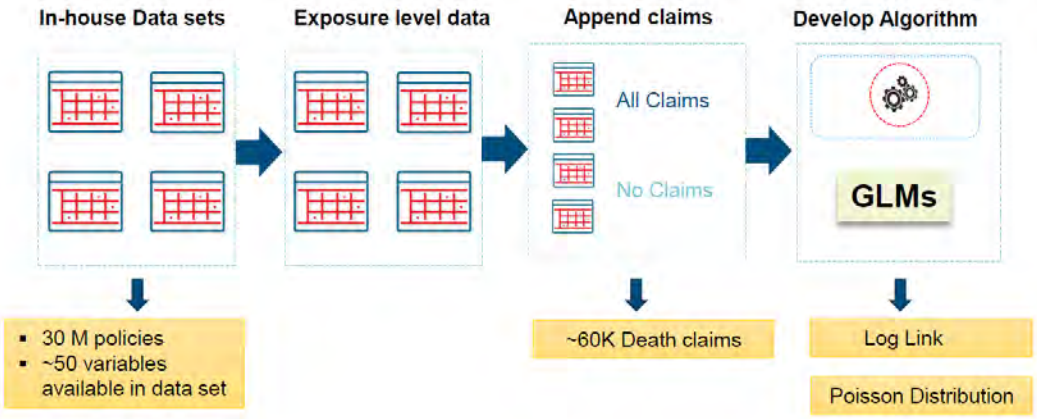
- GLMs helped the pricing teams to study the impact of non-traditional risk factors on mortality and compare it against the expected basis by adjusting for age and gender
- Enabled loadings at a more granular level due to better segregation of risks

Develop Predictive Models

- GLMs can be used to develop different types of predictive models to make predictions on claim frequencies, claim amounts, fraudulent claims etc.
- The final risk scores or risk categories can be used to provide customized offers to customers based on their overall risk profiles, simplify underwriting process , simplify the customer onboarding process etc.
- All datasets available in-house which contains policy level and claims information can be linked with other external third party data assets to develop models with high predictive power

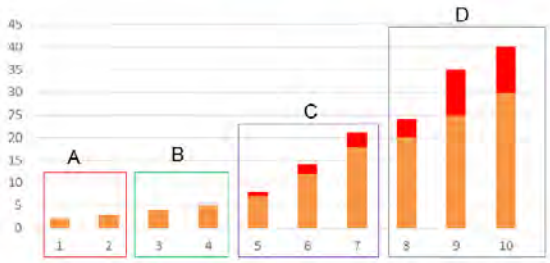
2 Case Studies

Case Study 1 – Claims Model



Case Study 1 – Claims Model

Lift Plot



Business Impact

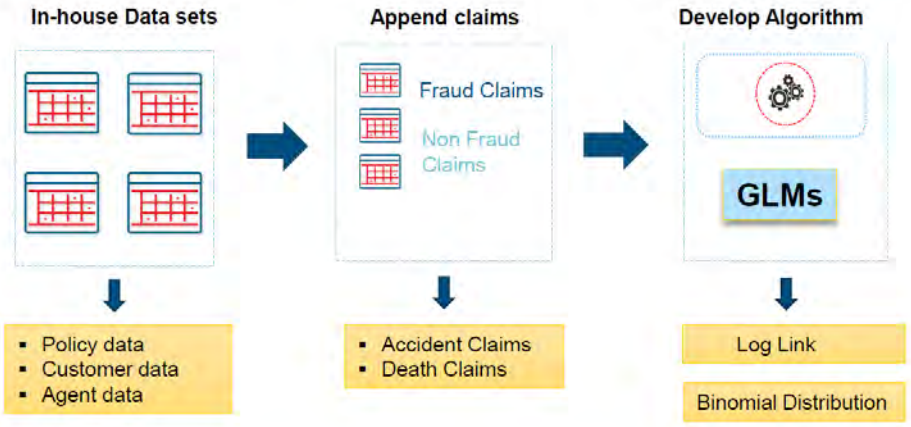
- Identify customers with very good risk profiles and provide offers to them on GIO / SIO basis
- Simplify underwriting to ease customer onboarding resulting in a boost to the sales volumes

Case Study 2 – Fraud Detection Model

Background

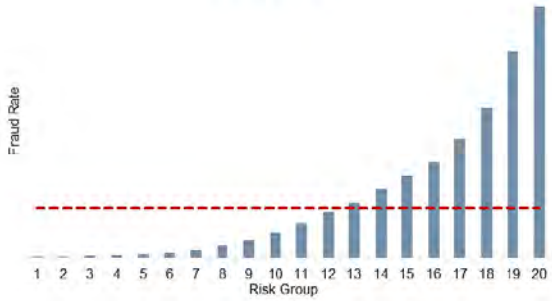
- Predict the likelihood of claim to be fraudulent at the stage of claim adjudication
- Increase capability to detect fraudulent claim with greater certainty and allow appropriate action to be taken
- Replace the existing rules based fraud detection engine with a advanced GLM based prediction model with additional features

Case Study 2 – Fraud Detection Model



Case Study 2 – Fraud Detection Model

Lift Plot



Business Impact

- Optimized deployment of claims investigation resources.
- No need to investigate the best groups, with more vigorous investigation for the worst groups.

Questions?

04/09/2019