8th Capacity Building Seminar
in Health and Care Insurance
15th March 2024

Predictive Loss Ratio using GLM and it’s applications

Speaker: Himanshu Manocha

Institute of Actuaries of India

Agenda

- Quick Recap on Generalized Linear Modelling (GLM)
- Predictive Loss Ratio in Health Insurance using GLM
  - Data Preparation and Considerations
  - Base distributions for Frequency and Severity Models
  - Selecting Rating Factors for Frequency and Severity Models
  - Checking interactions within the Rating Factors, if any
  - Tests to check the Fitment of both Models
  - Calculating Burn Cost or Risk Premium in order to arrive at the Predictive Loss Ratios
- Use of Predictive Loss Ratios in Portfolio Monitoring

Quick Recap on Generalized Linear Modelling (GLM) (1/2)

- GLM is a means of modeling the relationship between a variable whose outcome we wish to predict and one or more explanatory variables
- Few example of Predicted Variables are:
  - Claims Frequency
  - Claims Severity
  - Burn Cost / Pure Premium
  - Loss Ratio
  - Retention Ratio
  - Whether a submitted claim contains Fraud
- Explanatory variables are typically any Policy Terms or Policyholder or Claims Characteristics that you may wish to include in the model
- Goal in GLM modeling is to explain as much of the variability in the outcome as we can using explanatory variables
- The outcome of the Predicted Variable is assumed to be driven by both a Random Component and a Systematic Component

Quick Recap on Generalized Linear Modelling (GLM)
(2/2)

- <u>Random Component:</u>
  - Portion of the outcome driven by causes other than the explanatory variables in the model
  - Modelled as a random variable that follows a probability distribution, usually from an exponential family (which includes distributions such as Normal, Poisson, Gamma, Binomial, Tweedie)
  - Randomness of the outcome of a particular risk is expressed as:

Y ~ Exponential Family (μ, ϕ) (where μ is the mean and ϕ the dispersion parameter)

- <u>Systematic Component:</u>
  - Portion of the variation in outcome that is related to the explanatory variables
  - GLM models the relationship between "μ" (the model prediction) and explanatory variables as:

$$\underset{\substack{\Big\downarrow \\ \text{Link} \\ \text{Function}}}{G(\mu)} = \underbrace{\text{ß0} + \text{ß1X1} + \text{ß2X2} + \ldots.. + \text{ßpXp}}_{\substack{\Big\downarrow \\ \text{Linear} \\ \text{Predictor}}}$$

Predictive Loss Ratio in Health Insurance using GLM

Data Preparation and Considerations

Institute of Actuaries of India

- Policy Data:
  - Key Dataset as it serves as the foundation of Exposure Data which will be used in Frequency Model
  - This Dataset must be of the highest quality and completeness and should include key policy / policyholder related details such as:
    - ✓ Policy Start Date and End Dates
    - ✓ Member ID or Policy Number i.e. a unique ID that should enable linkage to the associated claims pertinent to that Member or Policy
    - ✓ Sum Insured
    - ✓ Location
    - ✓ Age
    - ✓ Gender
    - ✓ Distribution Channel

Policy No.	Member ID	Policy Start Date	Policy End Date	Age	Gender	Sum Insured
A	1	1/4/2022	31/3/2023	26	M	3,00,000
B	2	1/6/2022	31/5/2023	34	F	5,00,000
C	3	1/7/2022	30/6/2023	37	M	5,00,000
C	4	1/7/2022	30/6/2023	32	F	5,00,000
C	5	1/7/2022	30/6/2023	15	F	5,00,000
F	6	1/3/2023	29/2/2024	45	M	5,00,000
G	7	1/4/2023	31/3/2024	51	M	50,00,000
H	8	1/6/2023	31/5/2024	55	M	25,00,000
H	9	1/6/2023	31/5/2024	47	F	50,00,000
J	10	1/8/2023	31/7/2024	36	F	15,00,000
K	11	1/9/2023	31/8/2024	60	M	10,00,000

Predictive Loss Ratio in Health Insurance using GLM

Data Preparation and Considerations

Institute of Actuaries of India

- Claims Data:
  - This dataset will enable us to understand granularities within the claims which will be used to determine Frequency and Severity
  - Key technical and descriptive fields includes:
    - ✓ Claim Number
    - ✓ Policy Number and Member ID
    - ✓ Claim Amount
    - ✓ Type of Claim
    - ✓ Important Claim Dates like Date of Admission referred usually as the Loss Date

Policy No.	Member ID	Claim Number	Accident Date	Claim Type	Claim Amount
A	1	1	15/12/2022	In-Patient	1,50,000
B	2	2	19/9/2022	In-Patient	45,000
C	3	3	21/9/2022	In-Patient	65,000
C	4	4	6/8/2022	HC	5,000
G	7	5	15/1/2023	In-Patient	75,000
H	8	6	29/12/2023	In-Patient	90,000
H	8	7	31/1/2024	Post Hosp. Supplementary Claim	11,500
K	11	8	21/2/2024	In-Patient	1,05,000
F	6	9	20/8/2023	In-Patient	1,10,000

Predictive Loss Ratio in Health Insurance using GLM

Data Preparation and Considerations

Institute of Actuaries of India

- Other Considerations:
  - Combining Policy and Claims Datasets in order to get the key response variables such as Frequency and Severity together with other prospective Rating Factors
  - Segmentation of pricing data is critical in this exercise as it will define the risk buckets on which pricing will be based
  - Banding of existing variables from the Policy Data will help enrich this analysis as this approach allows for more accurate predictions of claims and setting premium which is reflective of the actual risk of each segment
  - Adjusting the Data – Decide on the treatment of supplementary or health check-up claims
  - Include data for the period with almost full maturity i.e. no significant further development is expected
  - Bifurcate Dataset into Test and Train Datasets i.e. Build GLM Model on Train Dataset and perform Actual vs Expected Test on Test Dataset

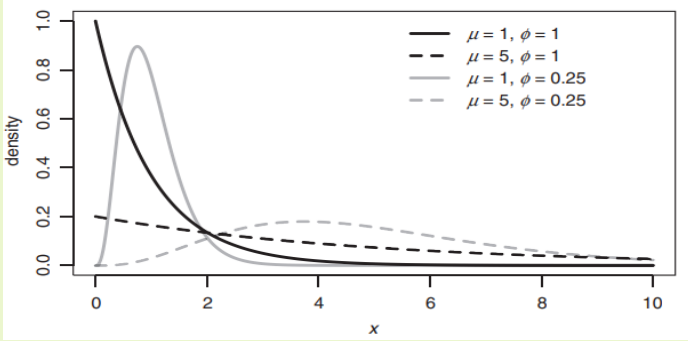
Predictive Loss Ratio in Health Insurance using GLM

Base Distributions for Frequency and Severity

Institute of Actuaries of India

- Commonly used Distributions for Frequency:
  - <u>Poisson Distribution</u>
    - ✓ Models the count of events occurring within a fixed time interval
    - ✓ Variance increases linearly with the Mean
    - ✓ Dispersion Parameter is 1
    - ✓ There may be scenarios where Overdispersion is observed in Claim Frequency i.e. variance is greater than the mean
    - ✓ To allow for this another source of variance i.e. variation in risk level among the policyholders one can use Overdispersed Poisson Distribution
  - <u>Negative Binomial Distribution</u> is another distribution to deal with overdispersion

- Commonly used Distributions for Severity:
  - <u>Gamma Distribution</u>
    - ✓ Right Skewed
    - ✓ Sharp Peak
    - ✓ Long tail to the right and has a lower bound to zero
    - ✓ Variance is proportional to an exponential function of mean
  - <u>Inverse Gamma Distribution:</u>
    - ✓ Similar characteristics as Gamma but with Sharper peak and wider tail
    - ✓ Appropriate where skewness of the severity is expected to be extreme



Predictive Loss Ratio in Health Insurance using GLM

Selecting Rating Factors or Explanatory Variables

Institute of Actuaries of India

Sample Rating Factors include:

- Sum Insured
- Gender
- Age
- Family Composition
- Distribution Channel
- Location
- Business Type
- Loss Year or Quarter to allow for the developments/trends in Predictive Loss Ratio

Variable Selection and Significance: With the selected factors, is the model able to explain as much variability as possible. Statistics to check for this:

- **Standard Error:** estimated standard deviation of the random process, the smaller the std. deviation the more confidence in the estimate
- **p-value:** estimate of the probability of a value of that magnitude or higher arising by pure chance. If the odd of a particular variable arising by pure chance is small, it is likely that the result reflects a real underlying effect – So the variable is significant

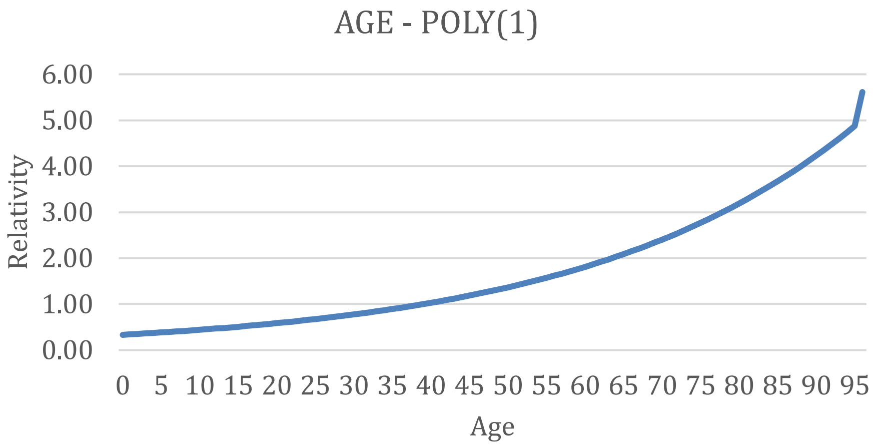
➢ Other Considerations or Checks:

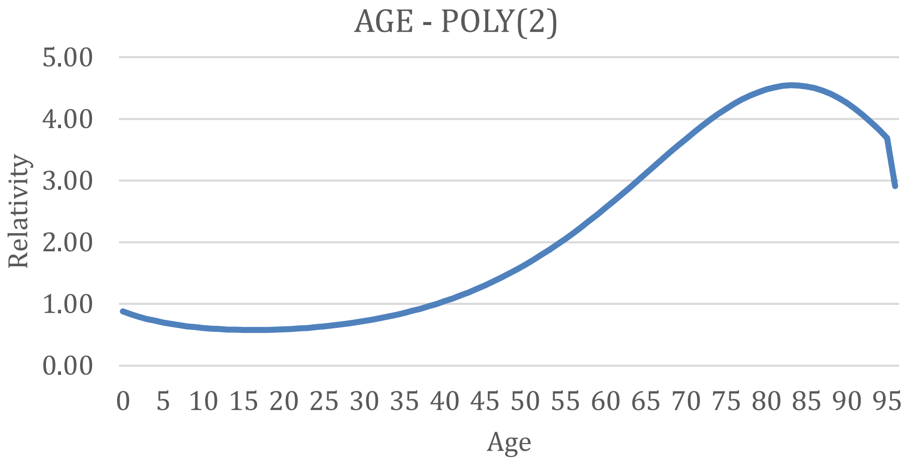
- Check for any Correlation or Multi-Collinearity among the factors
- Address Non Linearity, if any, present within the model (Illustration in the following slide)
- Check for Interactions within the model (Illustration in the following slide)
- Utilize Offsets if needed within the model

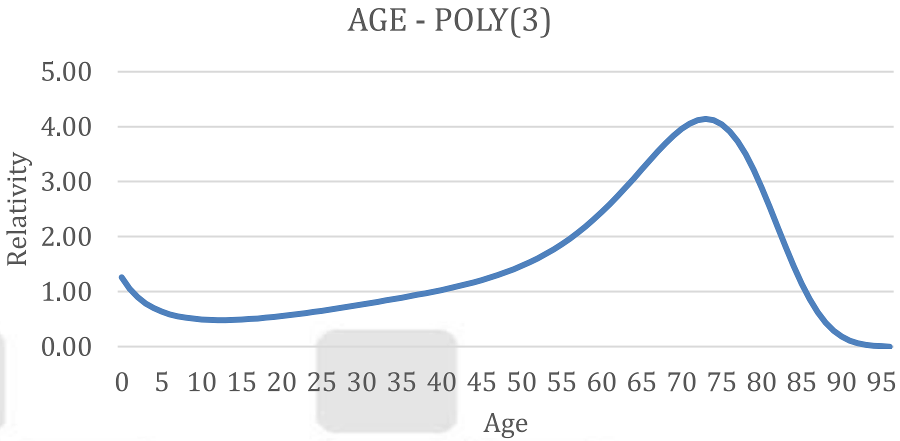
Predictive Loss Ratio in Health Insurance using GLM

Example to Address Non Linearity within Age Rating Factor

Institute of Actuaries of India

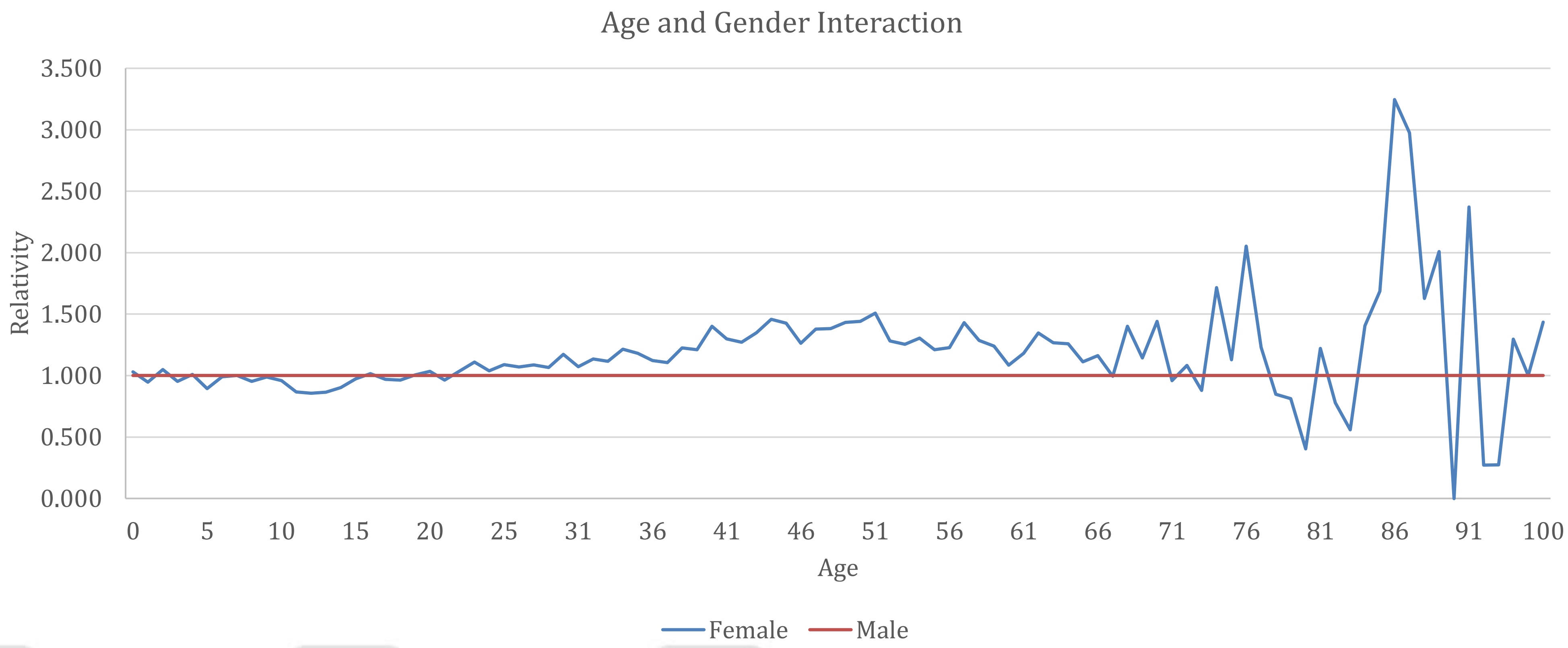






Predictive Loss Ratio in Health Insurance using GLM

Example for Interaction between Explanatory Variables



Predictive Loss Ratio in Health Insurance using GLM
Model Refinement

- Comparing Models:
  - Compare the deviance of two or more models – Did the added explanatory variables reduce the deviance significantly and at the same time maintaining or increasing the predictiveness of the model
  - F-Test or Chi-Square Test can also be done to compare the predictive power of two models
  - AIC can be used to compare models with different variables
- Checking the Fitment of Model and Selection of Model
  - Residual Analysis: Plotting residuals i.e. measures of deviation of actual from predicted
  - Check for Model Stability: Can be done by using Cook's Distance or bootstrapping technique
  - Assess fit with plots of Actual and Predicted
  - Measure Lift – to check model's ability to prevent adverse selection in other words to check for model's accuracy

Predictive Loss Ratio in Health Insurance using GLM

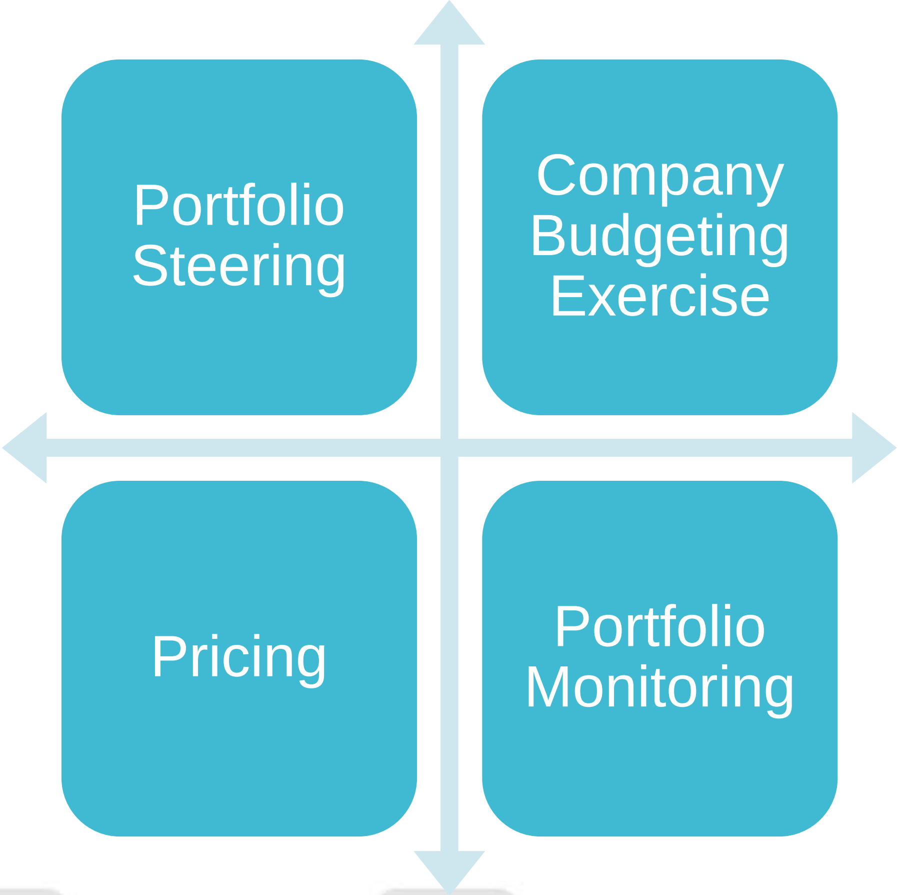
Illustration Showing Predictive Loss Ratio Calculation

Input Variables			
Model	Rating Factor	Input	Relativity
Frequency	Family Composition	One Adult	1.13
Frequency	Sum Insured	5,00,000	1.00
Frequency	Gender	Male	1.00
Frequency	Age	25	0.82
Severity	Sum Insured	5,00,000	1.04
Severity	Age	25	0.98

Base Numbers	
Base Frequency	6.50%
Base Severity	75,000

Calculation Steps	
Predicted Frequency	6.00%
Predicted Severity	76,440
Predicted Loss Amount	4,607
Actual Premium	12,000
Predicted Loss Ratio	**38.39%**

Applications of Predictive Loss Ratios



Portfolio Steering

Company Budgeting Exercise

Pricing

Portfolio Monitoring

Institute of Actuaries of India

Thank You!