SOCIETY OF ACTUARIES®

SOA Predictive Analytics Seminar – South Korea

30 Aug. 2019 | Seoul, South Korea

Session 2

The GLM and its application in general insurance

George Kau, ACMA, CGMA, FASM, FSA, MBA

Predictive Analytics

GLM and its application in general insurance

George Kau ACMA CGMA CA(M) FSA FASM MBA
KPMG PLT

August 2019

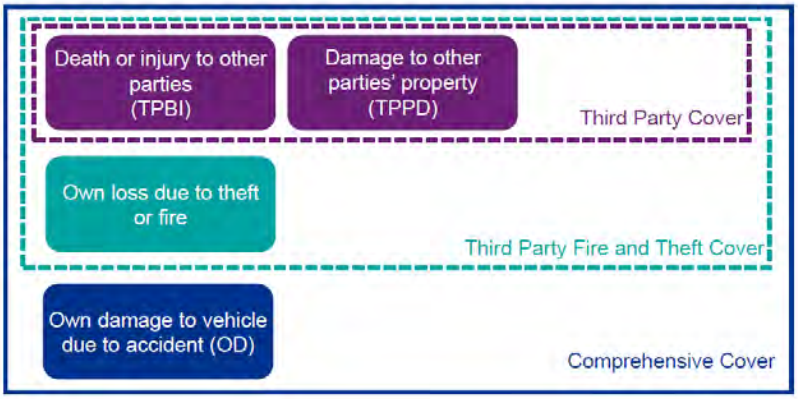
SOCIETY OF ACTUARIES

KPMG

Brief Introduction of Motor Insurance Rating in Malaysia

Motor Insurance
- Malaysia case study

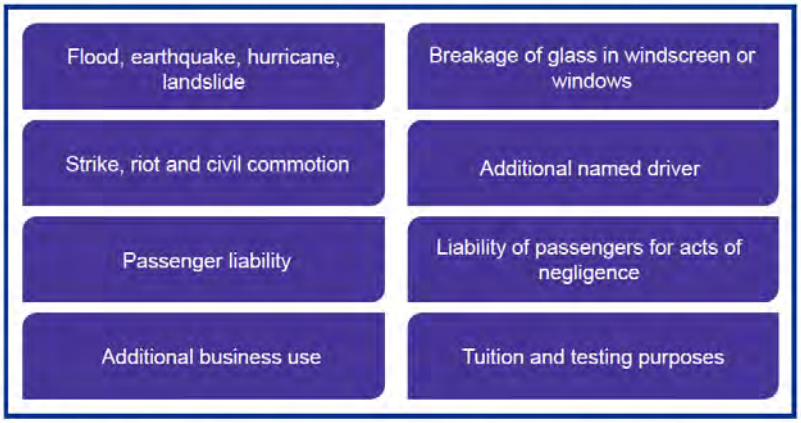
Motor Insurance in Malaysia is renewed yearly
Premiums are paid before insurance coverage starts



KPMG

Motor Insurance
- Extension Cover

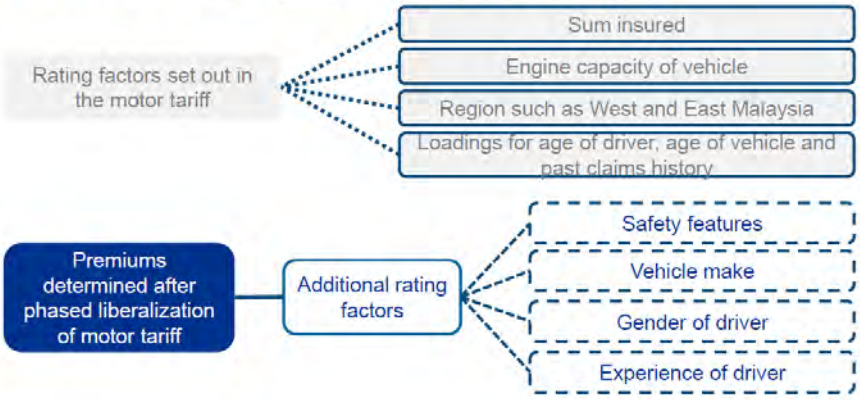
Additional perils can be added to the policy with additional premiums



KPMG

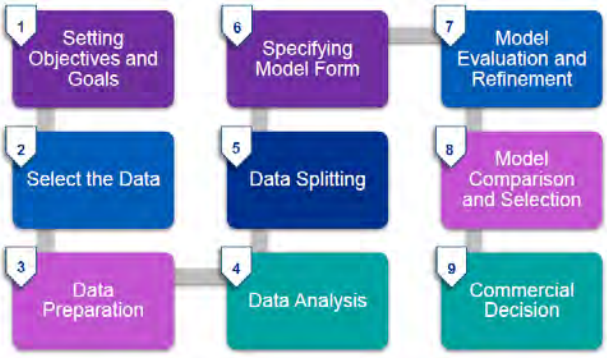
Motor Tariff
- Rating Factors

Premium rates charged by the insurance companies range within the allowable loading limit of Motor Tariff. It is a very competitive market for profitable growth.



Internal Pricing Process

A robust premium rating and governance process are required by the regulation.

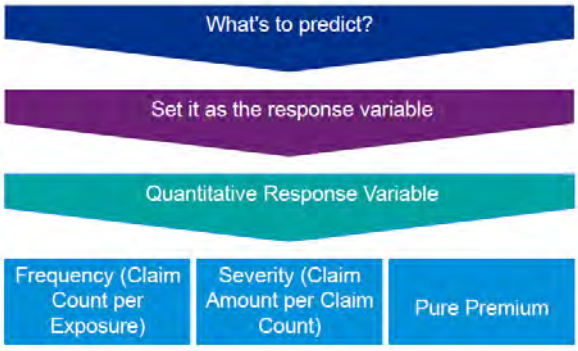


KPMG

GLM - Data Preparation Step 1 - 5

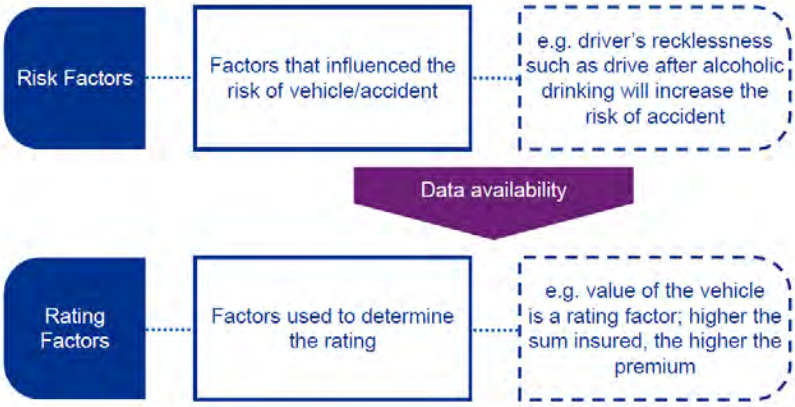
Step 1. Setting Objectives and Goals

- Purpose of Modelling



Step 2. Select the Data

- Risk Factor Vs. Rating Factor



Step 2. Select the Data (cont'd)

- Driver Factors Category

Rating Factor	Description	Data Structure
Age of Driver	Age of **vehicle owner**, or age of policyholder	Integer
Driving Experience	**Length** of **driving period** or Experience	Integer
Driving Record	**Number** of **traffic offences** or **bad record**	Integer
Gender	Male or Female	Categorical
Marital Status	Single or Married	Categorical
Number of Driver	**List of drivers** in the policy	Integer

Step 2. Select the Data (cont'd)

- Vehicle Factors Category

Rating Factor	Description	Data Structure
Cubic Capacity	**Dimension of vehicle engine**	Integer
Manufactured Year	Number of **years** since the vehicle is manufactured	Integer
Safety Features	Number of **safety installations**	Integer
Odometer	**Distance travelled** by the vehicle	Numerical
Vehicle Type	Sports or Normal vehicle	Categorical

Step 2. Select the Data (cont'd)

- Location Factors Category

Rating Factor	Description	Data Structure
Region	**East** or **West Malaysia**	Categorical
Address Location	**Postcode**	Categorical
Urbanization Level	**City, rural and suburban**	Categorical

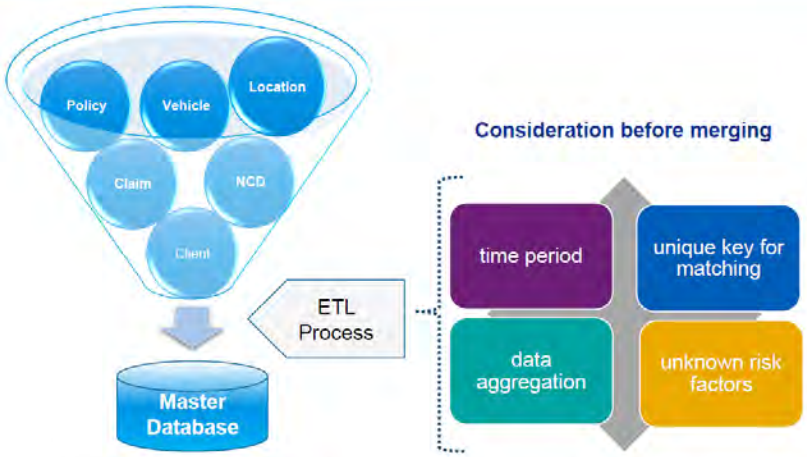
Step 2. Select the Data (cont'd)

- Policy Factors Category

Rating Factor	Description	Data Structure
Sum Insured	Market value or agreed value of the vehicle	Numerical
Policy Coverage	**Type** of coverage	Categorical
Renewal Indicator	**New** business or **renewal** Business	Categorical
Claim Count Experience	**Number** of claim incurred in the past	Integer
Claim Amount Experience	**Amount** of claim incurred in the past	Numerical
No Claim Discount (NCD)	Discount offered for good driving record	Numerical

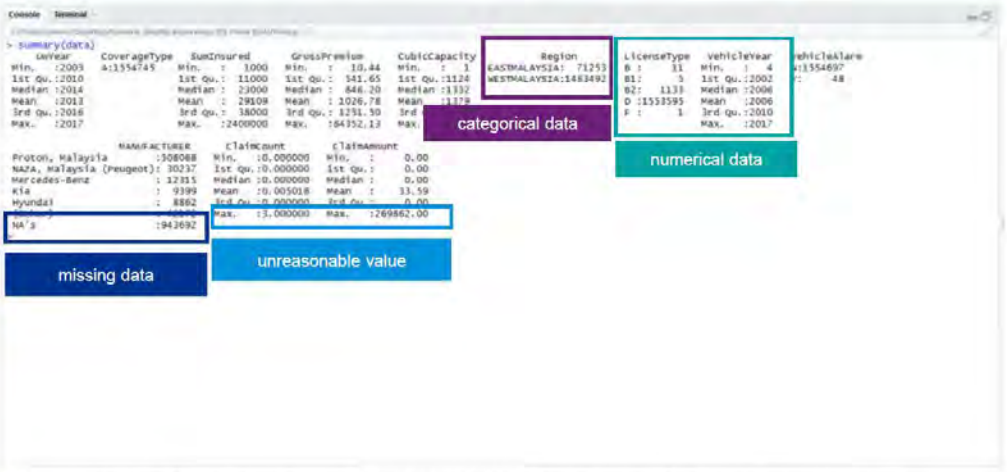
Step 3. Data Preparation

- Merging and Consideration



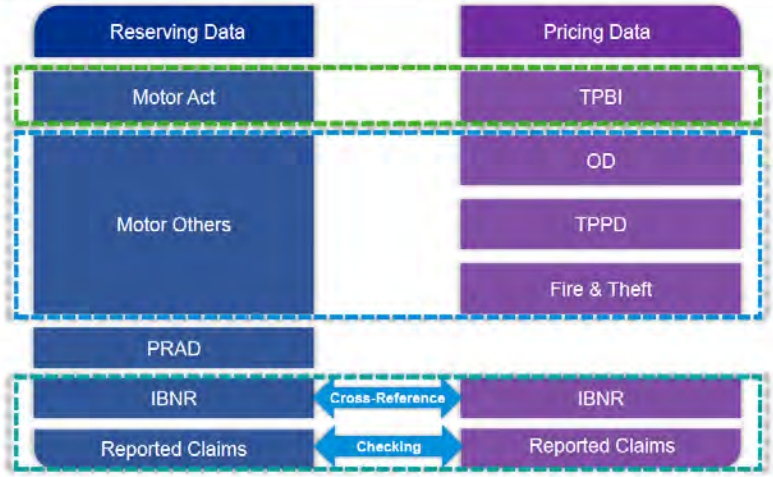
Step 3. Data Preparation (cont'd)

- Merging and Consideration



Step 4. Data Analysis

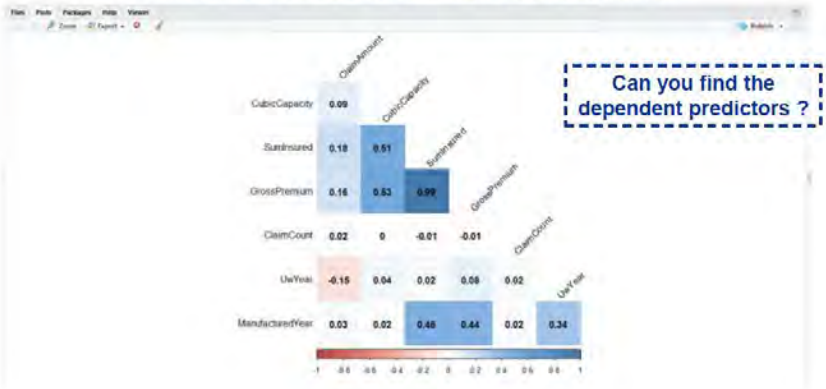
- Reserving vs Rating



Step 4. Data Analysis (cont'd)

- Correlation Plot

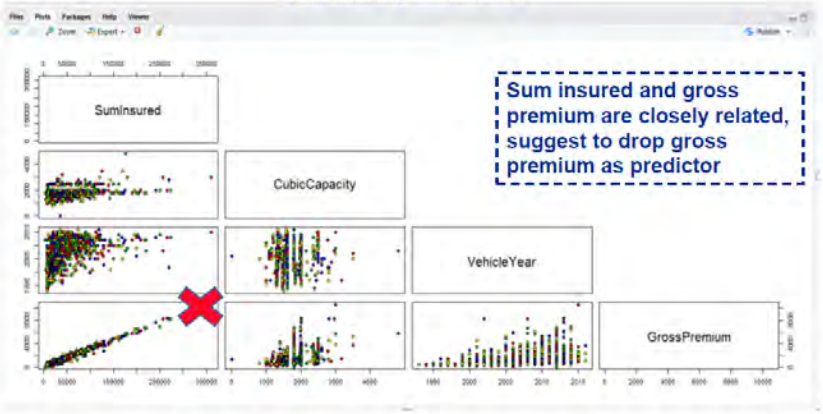
Correlation Plot – Pearson Coefficient Correlation Method



Step 4. Data Analysis (cont'd)

- Relationship Pattern Plot

Relationship Pattern Plot



Step 5. Data Splitting
- Training and Validation Sets

Training Set	(*70%*) to **BUILD** the GLM model using rating factors
Test Set	(*30%*) to **REFINE** the GLM model

KPMG

GLM - Modelling
Step 6 - 9

Generalized Linear Model
- Response variable

Regression analysis is a form predictive modeling technique which investigates the **relationship** between a **response variable** Y and the **predictors** X_i

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon$$

Response variable

Specifies the explanatory variables $(X_1, X_2, \ldots X_N)$ in the model

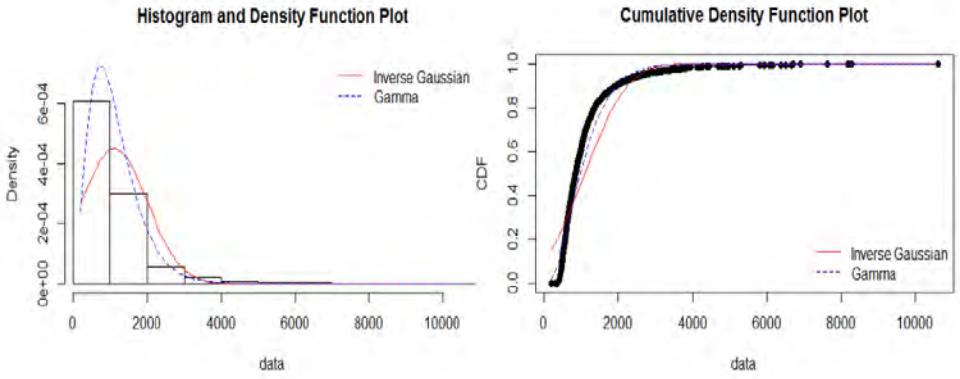
Generalized Linear Model (cont'd)
- Response variable

Continuous Response Variables e.g. severity, net premium	Inverse Gaussian / Gamma Regression
Count Response Variables e.g. claim count	Poisson / Negative Binomial Regression
Categorical Response Variables e.g. fraud, lapse (yes or no)	Binomial / Logistic Regression

Generalized Linear Model (cont'd)

- Response variable

Gamma distribution v.s. Inverse Gaussian distribution for Severity Model



Generalized Linear Model (cont'd)

- Response variable

Distribution	Typical Uses	Support of Distribution
Gaussian (Normal)	Linear response data, constant increments or decrements	Real: $(-\infty, +\infty)$
Inverse Gaussian	Positively skewed data with distribution's tail decreases slowly	Real: $(0, +\infty)$
Gamma	Exponential response data, increase or decrease with constant ratio	Real: $(0, +\infty)$

Distribution	Typical Uses	Support of Distribution
Binomial	Single outcome from N occurrences	Integer: 0,1,2 ..., N
Poisson	Count data	Integer: 0,1,2 ...

Generalized Linear Model (cont'd)

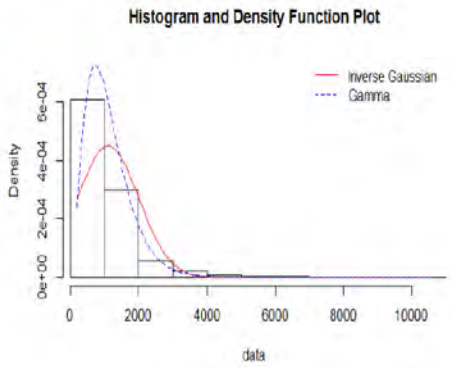
- Link Function

The relationship between the mean of the response variable distribution function and a linear combination set of predictors

Numerical example for a Gamma Log Link Model

$$\ln[E(Severity_i)] = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$
$$E(Severity_i) = Exp\{\alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p\}$$

$$\mu_i = E[Y_i] = E(Severity_i) = 3{,}000$$
$$\ln[E(Severity_i)] = \ln[3{,}000] = 8.01$$
$$E(Severity_i) = Exp\{8.01\} = 3{,}000$$



Generalized Linear Model (cont'd)

- Link Function

Distribution	Link Name	Link Function, $X\beta = g(\mu)$	Mean Function
Normal	Identity	$X\beta = \mu$	$\mu = X\beta$
Inverse Gaussian	Inverse Squared	$X\beta = \frac{1}{\mu^2}$	$\mu = (X\beta)^{-\frac{1}{2}}$
	Log	$X\beta = \ln(\mu)$	
Gamma	Inverse	$X\beta = \frac{1}{\mu}$	$\mu = -(X\beta)^{-1}$
	Log	$X\beta = \ln(\mu)$	
Binomial	Logit	$X\beta = \ln(\frac{\mu}{1-\mu})$	$\mu = \frac{\exp(X\beta)}{1+\exp(X\beta)}$
Poisson	Log	$X\beta = \ln(\mu)$	$\mu = \exp(X\beta)$

Generalized Linear Model (cont'd)

- Weights

In insurance, the greater risks should carry more weight in the estimation of the model coefficients

For example, the exposure for each observation is different, one observation might relate to one month's exposure, and another to one year's exposure.
The analysis weights are "known" value that can vary from observation to observation.

Numerical example for a Gamma Log Link Model

$$\ln[E(Severity_i)] = \alpha + w_1X_1 + w_2X_2 + \cdots + w_pX_p$$

$$E(Severity_i) = Exp\{\alpha + w_1X_1 + w_2X_2 + \cdots + w_pX_p\}$$

where w is the vector of weighted model coefficients

Generalized Linear Model (cont'd)

- Offsets

When the effect of predictor is known, it is appropriates to include the information into the model as a known effect.

This can be achieved by using the "*offset term*", whose coefficient is constrained to be 1.

Numerical example for a Gamma Log Link Model

$$\ln[E(Severity_i)] = \alpha + \beta_1X_1 + \beta_2X_2 + \cdots + \beta_pX_p + \ln(offset)$$

$$E(Severity_i) = Exp\{\alpha + \beta_1X_1 + \beta_2X_2 + \cdots + \beta_pX_p\} \times offset$$

Step 6. Specifying Model Form

- Severity Model Example

Training Set

Objective	Predict the Expected Severity of Motor Insurance
Response Variable	Severity = Claim Amount / Claim Count
Predictors	Sum Insured, Underwriting Year, Cubic Capacity of Vehicle, Manufacturer of Vehicle, Manufactured Year, Region
Weights	Claim Count
Models	Inverse Gaussian Distribution or Gamma Distribution
Link Function	Log Link Inverse Gaussian and Log Link Gamma

Step 7. Model Evaluation and Refinement

- Goodness of Fit Test

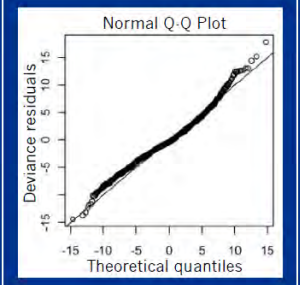
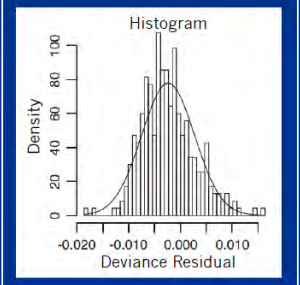
Training Set

Measurements	Conditions / Formula
Log-Likelihood	$ll_{null} \leq ll_{model} \leq ll_{saturated}$
Deviance	$Deviance = 2 \times (ll_{saturated} - ll_{null})$
Akaike Information Criterion (AIC)	$AIC = -2(log\text{-}likelihood) + 2(n_{paramter})$
Akaike Information Criterion Corrected (AICc)	$AICc = AIC + \frac{2k^2 + 2k}{n - k - 1}$
Bayesian Information Criterion (BIC)	$AIC = -2(log\text{-}likelihood) + k \times log(n)$

Step 7. Model Evaluation and Refinement

- Goodness of Fit Test (Cont'd)

Training Set

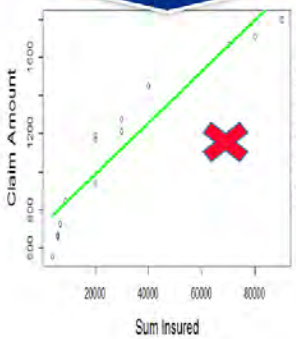
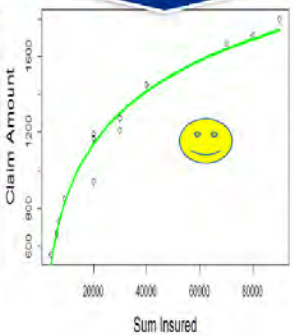
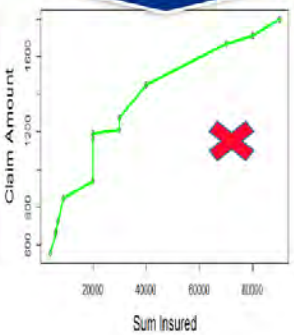
Measurement	q-q Plot	Deviance Residual (Histogram)
Conditions / Formula		

Step 7. Model Evaluation and Refinement

- Other Evaluation Criteria

Training Set

Model Validation — Test for over-fitting or under-fitting using validation set

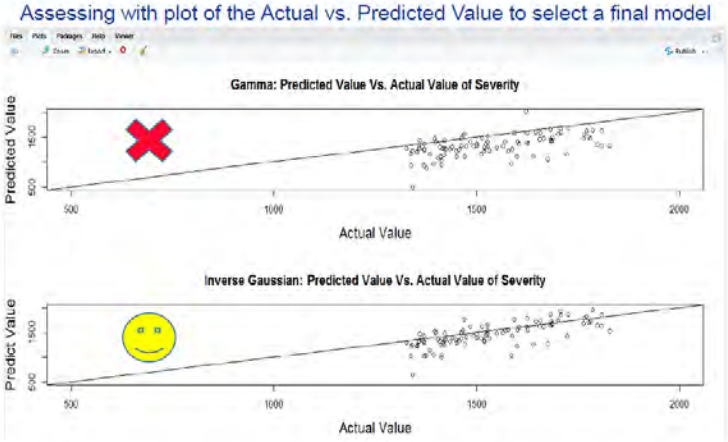
Under-fitting	Fitting	Over-fitting
		

Step 8. Models Comparison and Selection

- Comparing Models

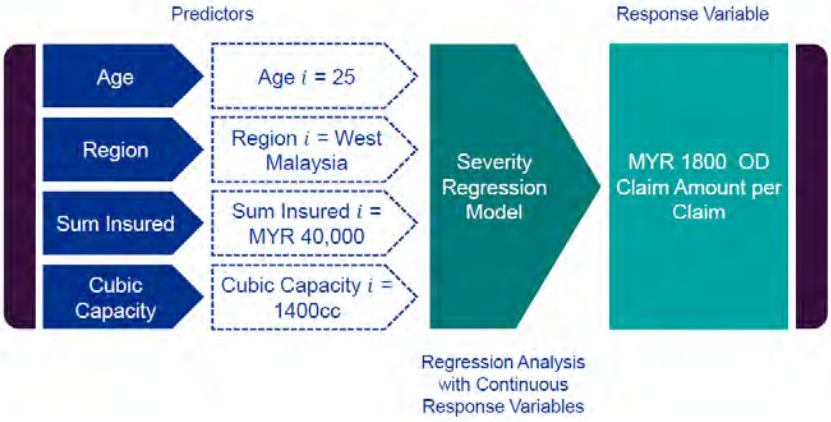
Test Set

Assessing with plot of the Actual vs. Predicted Value to select a final model



Step 8. Models Comparison and Selection

- Final Model



Step 8. Models Comparison and Selection

- Final Model (cont'd)

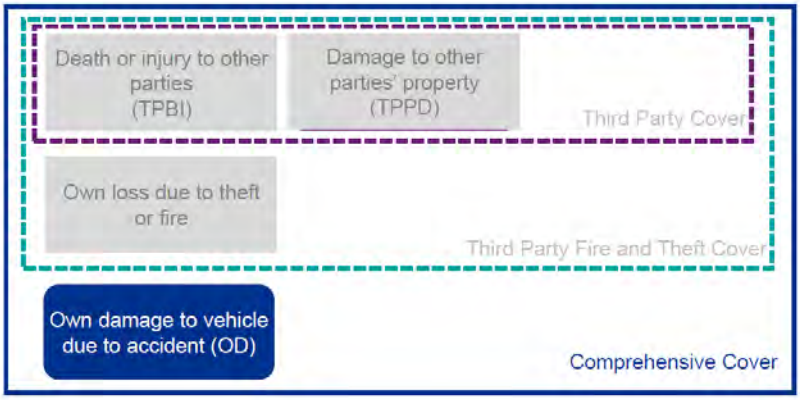
Any trending adjustments will take place at the frequency and severity model level (judgement required)

OD Excess is the estimated loading for the large losses excluded from the dataset (judgement required)

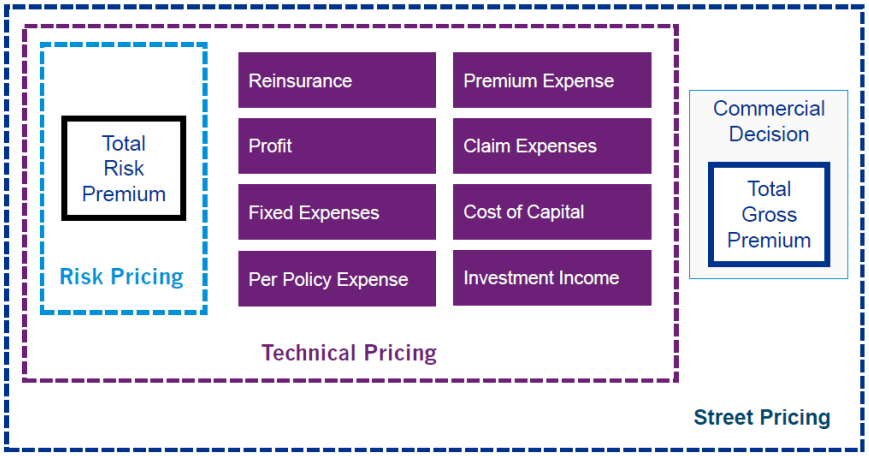
Step 8. Models Comparison and Selection

- Net Rating

Step 9. Commercial Decision
- Malaysia case study



Step 9. Commercial Decision
- Gross Rating



KPMG

Thank you

Questions

KPMG

kpmg.com/socialmedia

Download on the App Store GET IT ON Google Play

kpmg.com/app

SOCIETY OF ACTUARIES