Use of GLMs in a competitive market

Ji Yao and Simon Yeung, Advanced Pricing Techniques (APT) GIRO Working Party

12 June 2013
About the presenters

Dr. Ji Yao is a manager with Ernst & Young’s EMEIA insurance risk and actuarial services practice. He has extensive first-hand experience in various modelling for pricing, including risk models, demand models and price optimisation, with a solid background in mathematics and statistics. He is the chair of the Advanced Pricing Techniques (APT) GIRO working party.

JYao@uk.ey.com

Simon Yeung is currently a senior manager at Grant Thornton. Prior to joining Grant Thornton, Simon was the head of motor pricing at Saga for 3 years. Before that, he was a reserving manager at RBS Insurance for 3 years. Before joining RBS Insurance he worked for London market insurers, reinsurers and commercial insurers for four and half years. He is a member of the Advanced Pricing Techniques (APT) GIRO working party.

Simon.yeung@uk.gt.com
Agenda

• Introduction

• Current market and uses of GLM

• Three overlooked facts of GLM and their implications

• Summary and Q&A
Introduction

• Advanced Pricing Techniques (APT) GIRO working party was created in 2012
• 22 members working in three work streams
  – GLM
  – Telematics pricing
  – Conversion/Elasticity modelling
• One workshop in GIRO 40 and one paper on GLM is being prepared
• We will focus on GLM in this presentation
Current uses of GLM in the market

- Risk base pricing
- Cost plus approach
- Price optimisation
Market Performance

Looking in more detail at:

• Change in claim ratio and frequency over time
• What, if any, relationships can we derive between the two?
• How does this relate back to GLM modelling?

Data used:

• Cross section of market (8 companies)
• Totalling £4.7bn earned premium in 2010
• High level data taken from FSA returns
Steady rise in claim ratio from 2005 due to reasons such as:

- aggregators
- increase in BI claims
- recession
- claims farming

Fall in claims ratio 2009-2010 due to rate increase
Some companies have seen their claims frequency fluctuate greatly over time. On average claim frequency has been slowly reducing since 2007.
Correlation between Frequency and ULR

Average Claims Ratio vs Average Claims Frequency

Strong correlation (~74%) between Claims Ratio and Claims Frequency

- Market competition putting pressure on price can charge
- Insufficient data to accurately price the risks

2008-2010 excluded due to recession and hike in petrol prices

- 1. Company data is only a sample
- 2. Unable to model using GLMs

Year

Average Claims Frequency [%]

Average Claims Ratio
Frequency vs Average Premium

![Graph showing the relationship between claims frequency and average premium over years. The graph has two lines: one representing average claims frequency excluding companies 2 and 8, and the other representing average premium excluding companies 2 and 8. The x-axis represents the years from 2001 to 2007, and the y-axes represent claims frequency and average premium. The graph shows a decrease in claims frequency and an increase in average premium over the years.]
Change in mix of business

- Fluctuations in frequency due to changes in mix of business

![Claims Frequency vs Claims Ratio (Company 2)](image1)

![Claims Frequency vs Claims Ratio (Company 8)](image2)
Claim Frequency vs Average Premium over time (Company 2)

Claim Frequency vs Average Premium over time (Company 8)
Quotes for a 30 year old male with a clean license held for 10 years, for a 57 plate manual 1.6L ford focus style 5 door hatchback. Car is kept at home parked on the road, for social use only, approx 9000 annual mileage.
<table>
<thead>
<tr>
<th>Provider</th>
<th>Annual Premium</th>
<th>Monthly Premium</th>
<th>Excess</th>
<th>Legal Cover</th>
<th>Courtesy Car</th>
<th>Breakdown Cover</th>
</tr>
</thead>
<tbody>
<tr>
<td>Esure</td>
<td>£200.09</td>
<td>£16.67</td>
<td>£17.94</td>
<td>£17.94</td>
<td>✔</td>
<td></td>
</tr>
<tr>
<td>M&amp;S Bank</td>
<td>£259.73</td>
<td>£21.64</td>
<td>£22.27</td>
<td>£22.27</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Post Office</td>
<td>£294.23</td>
<td>£24.52</td>
<td>£28.93</td>
<td>£28.93</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Halifax</td>
<td>£322.72</td>
<td>£26.89</td>
<td>£29.55</td>
<td>£29.55</td>
<td>✔</td>
<td></td>
</tr>
<tr>
<td>ASDA Money</td>
<td>£323.46</td>
<td>£26.95</td>
<td>£31.31</td>
<td>£31.31</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>£376.59</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Quotes for a 40 year old married female with a clean license held for 15 years for a 59 diesel Golf GTD 2.0L 3 door hatchback. Car is kept at home and parked on a driveway for social use only, approx 7000 miles.
Current market and uses of GLM

• GLM is a standard approach for risk pricing and price optimisation
• Wide range of price for individual quote
• Wide range of performance for market player
• What causes the difference?
GLM technical details

A GLM consists of the following three components:

1. **Random component**
   Each component of $Y$ is independent and is from one of the exponential family of distributions.

2. **Systematic component**
   A linear combination of the estimated parameters gives the linear predictor, $\eta$:
   \[ \eta = X \beta \]

3. **Link function**
   The relationships between the random and systematic components is specified via a link function, $g$, such that:
   \[ E[Y] \equiv \mu = g^{-1}(\eta) \]

4. **Data**
   The dataset that GLM trained on.
Three overlooked facts of GLM

1. GLMs put either zero or full credibility into data
2. GLMs implicitly use median from the distribution of prediction
3. GLM results depend on the mixture of rating variables in the data
Quiz 1: Average weight of yellow balls

There is a bag of coloured balls. You sampled a few of them from the bag and obtained the following information:

<table>
<thead>
<tr>
<th>Colour</th>
<th>Avg weight (kg)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yellow</td>
<td>6</td>
</tr>
<tr>
<td>Red</td>
<td>4</td>
</tr>
<tr>
<td>Average All</td>
<td>5</td>
</tr>
</tbody>
</table>

What is your estimation of the average weight of yellow balls?

A) Use average of yellow balls ONLY – 6kg
B) Use average of ALL balls – 5kg
C) Blended average weight of yellow balls and non-yellow balls
D) Other (with suggestions)
GLM fact 1: GLMs put either zero or full credibility into data

\[
\eta = X \cdot \beta \\
\downarrow \\
E(Y) \equiv \mu = g^{-1}(\eta)
\]
A gradual approach to include data is needed in modelling

Sample 6 balls from the bag of yellow and red balls, and we obtained these weights:

<table>
<thead>
<tr>
<th>Colour</th>
<th>Avg weight (kg)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yellow</td>
<td>4</td>
</tr>
<tr>
<td>Yellow</td>
<td>6</td>
</tr>
<tr>
<td>Yellow</td>
<td>8</td>
</tr>
<tr>
<td>Red</td>
<td>2</td>
</tr>
<tr>
<td>Red</td>
<td>4</td>
</tr>
<tr>
<td>Red</td>
<td>6</td>
</tr>
</tbody>
</table>

Testing the ‘colour’ factor in a GLM shows that Yellow is not significantly different from Red at 95% confidence level (p-value=0.1336).

Avg weight of yellow balls = 5

Keep sampling and if we get 6 more identical balls as before:

<table>
<thead>
<tr>
<th>Colour</th>
<th>Avg weight (kg)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yellow</td>
<td>4</td>
</tr>
<tr>
<td>Yellow</td>
<td>4</td>
</tr>
<tr>
<td>Yellow</td>
<td>6</td>
</tr>
<tr>
<td>Yellow</td>
<td>6</td>
</tr>
<tr>
<td>Yellow</td>
<td>8</td>
</tr>
<tr>
<td>Yellow</td>
<td>8</td>
</tr>
<tr>
<td>Red</td>
<td>2</td>
</tr>
<tr>
<td>Red</td>
<td>2</td>
</tr>
<tr>
<td>Red</td>
<td>4</td>
</tr>
<tr>
<td>Red</td>
<td>4</td>
</tr>
<tr>
<td>Red</td>
<td>6</td>
</tr>
</tbody>
</table>

Testing the ‘colour’ factor in a GLM shows Yellow is now significantly different from Red at 95% confidence level (p-value=0.0339).

Avg weight of yellow balls = 6

Would you suddenly change your view because of the additional six balls?
An important implication is GLMs tend to push relativities and hence price towards extreme levels

• As the normal GLM practice is to calibrate the base rate after relativities are calculated, extreme relativities will result in more policies being priced at very low (or high) end

• Over-priced policies never get converted in a competitive market, so insurers are exposed to big under-pricing risk

• Linked to the observed diversified quoted premium on the market
Generalised linear mixed models (GLMMs) provide a potential solution

• GLMMs are an extension to GLM, in which the linear predictor contains random effects to allow for correlation of the data in addition to the usual fixed effects.

• It provides a convenient way of applying credibility blending within GLM.

\[ y_i = X_i \beta + Z_i b_i + \epsilon_i. \]

Random effect
Quiz 2: Mean, median or mode? – a question not only relevant to reserving or capital

- A pricing analysis gives a range of possible prices for a risk as shown in the table below:

<table>
<thead>
<tr>
<th>Price</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>£400</td>
<td>20%</td>
</tr>
<tr>
<td>£500</td>
<td>30%</td>
</tr>
<tr>
<td>£600</td>
<td>20%</td>
</tr>
<tr>
<td>£700</td>
<td>20%</td>
</tr>
<tr>
<td>£800</td>
<td>10%</td>
</tr>
</tbody>
</table>

- What is the price you will charge for the risk?

A) Mode - £500  
B) Median - £550  
C) Mean - £560  
D) Other (with suggestions)
GLM fact 2: GLMs implicitly use median from the distribution of prediction

- The linear predictor $\Sigma X_i \beta_i$ is asymptotically normally distributed as all $\beta_i$ are asymptotically multivariately normally distributed.
- After the link function transformation, the prediction is no longer normally distributed.
- Take log link as an example:

$$e^{X^T \beta} = e^{\Sigma X_i \beta_i} \quad \text{where} \quad \beta_i \sim N(\beta_i, \sigma^2)$$

Mean:
$$E[e^{\Sigma X_i \beta_i}] = e^{\Sigma \mu_i + \frac{1}{2} \text{var}(\Sigma X_i \beta_i)}$$

Median:
$$E[\Sigma X_i \beta_i]$$

Mode:
$$e^{\Sigma X_i \beta_i - \text{var}(\Sigma X_i \beta_i)}$$
Link function is the dominant factor in shaping the distribution of prediction

Consider a severity model with Gamma error structure. Results for different link functions:

These examples show that the upper and lower bounds could be very different, and the prediction is not always the mean of the distribution!
Do GLMs systematically underestimate the cost?

- For a distribution skewed towards the left, usually it is the case that Mode < Median < Mean, so the median used by GLMs is always lower than the mean.

- To use mean, the term $\var{\sum X_i/\beta_i}$ needs to be better understood and calculated. The key difficulty is the correlation matrix between $\beta_i$. 

---

Data

\[ X \]

\[ \eta = X \cdot \beta \]

\[ E(Y) = \mu = e^{-\eta} \]
GLM fact 3: GLM results depend on the mixture of rating variables in the data

<table>
<thead>
<tr>
<th>Driver Age</th>
<th>Car Age</th>
<th>Claim</th>
</tr>
</thead>
<tbody>
<tr>
<td>Old</td>
<td>Old</td>
<td>0.2</td>
</tr>
<tr>
<td>Old</td>
<td>New</td>
<td>0.3</td>
</tr>
<tr>
<td>Young</td>
<td>Old</td>
<td>0.4</td>
</tr>
<tr>
<td>Young</td>
<td>New</td>
<td>0.6</td>
</tr>
</tbody>
</table>

Parameter | Level1 | Estimate | StdErr |
---|--------|----------|--------|
Intercept |        | 0.4286   | 0.565  |
age Old    |        | -0.2411  | 0.6061 |
age Young  |        | 0        | 0      |
carage New |        | 0.1339   | 0.5836 |
carage Old |        | 0        | 0      |
Scale      |        | 1        | 0      |

Parameter | Level1 | Estimate | StdErr |
---|--------|----------|--------|
Intercept |        | 0.4305   | 0.5594 |
age Old    |        | -0.2374  | 0.5827 |
age Young  |        | 0        | 0      |
carage New |        | 0.1297   | 0.552  |
carage Old |        | 0        | 0      |
Scale      |        | 1        | 0      |
GLMs results are dragged toward the segment where there is more data

4 data points

<table>
<thead>
<tr>
<th>Driver Age</th>
<th>Car Age</th>
<th>Claim</th>
<th>Prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Old</td>
<td>Old</td>
<td>0.2</td>
<td>0.1875</td>
</tr>
<tr>
<td>Old</td>
<td>New</td>
<td>0.3</td>
<td>0.32143</td>
</tr>
<tr>
<td>Young</td>
<td>Old</td>
<td>0.4</td>
<td>0.42857</td>
</tr>
<tr>
<td>Young</td>
<td>New</td>
<td>0.6</td>
<td>0.5625</td>
</tr>
</tbody>
</table>

5 data points

<table>
<thead>
<tr>
<th>Driver Age</th>
<th>Car Age</th>
<th>Claim</th>
<th>Prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Old</td>
<td>Old</td>
<td>0.2</td>
<td>0.19315</td>
</tr>
<tr>
<td>Old</td>
<td>Old</td>
<td>0.2</td>
<td>0.19315</td>
</tr>
<tr>
<td>Old</td>
<td>New</td>
<td>0.3</td>
<td>0.3229</td>
</tr>
<tr>
<td>Young</td>
<td>Old</td>
<td>0.4</td>
<td>0.43053</td>
</tr>
<tr>
<td>Young</td>
<td>New</td>
<td>0.6</td>
<td>0.56027</td>
</tr>
</tbody>
</table>

• The dependency is not trivial. Some practical examples are:
  • Quote based premium model vs. sale based premium model
  • Modelled loss ratio for quotes vs. Sales
  • Time testing
With a view to future is the key to mitigate this issue

- GLM should be trained on expected future mixture of portfolio, rather than historical portfolio.

- Iterative modelling approach:

  Fit GLM → Set price
  Feed weight into GLM ↔ Model conversion

Data

\[ \eta = X \cdot \beta \]
\[ E(Y) \equiv \mu = g^{-1}(\eta) \]
Summary

• Significant variation in underwriting performance and quoted premiums in the current motor market pose challenges on the pricing techniques used in business.

• As the standard pricing technique, GLMs are coming cross new issues in a highly competitive market:
  - GLMs put either zero or full credibility into data
  - GLMs implicitly use median from the distribution of prediction
  - GLM results depend on the mixture of rating variables in the data

• Being able to understand and solve these issues could be one of the key ways to gain a competitive advantage in the market.
Expressions of individual views by members of the Institute and Faculty of Actuaries and its staff are encouraged.

The views expressed in this presentation are those of the presenter.