Actuarial Data Analytics / Predictive Modeling Seminar Part 1

Wifi details:
ILTCI Conference
ILTCI2016
Actuarial

Data Analytics / Predictive Modeling Seminar Part 1

Jim Berger
Matt Morton
Ben Williams
Paul Bailey

16th Annual Intercompany Long Term Care Insurance Conference
The data used in this seminar is fictitious, and it or results of analyses based on it should not be relied upon for any purpose.
1. Background to the Workshop
2. Predictive Modeling and Applications to Long Term Care Insurance
3. Predictive Modeling of Long Term Care Incidence using Generalized Linear Models in Emblem
Agenda of today’s session

- Towers Watson Emblem
  - What is it and why are we using it for the workshop?
  - Ensuring everyone has access to it
- What predictive modeling data looks like
- Preliminary Analyses in Emblem using LTC incidence data
  - Correlation analysis
  - Univariate/bivariate analysis
  - Distribution of response (discussion only)
- Statistical Background to Generalized Linear Models (GLMs)
Logging on to Microsoft Azure

Open The Azure Remote App:

Login:  TrainXX@wtwsaas.com
Password:  T0wersW@tson16
where XX is the number written on your handout
• What is it?
  – Market-leading GLM-fitting software

• Why are we using it in this seminar?
  – It is easy to use (no programming required), fast and visual

• Ensuring everyone has access to it:
  – Log on to Azure, taking note of your username and password
  – If you can open Emblem and get to a screen like the one shown, you are good to go
What is Predictive Modeling?

• We have a series of historical observations showing how input variables \(x\) (independent variables) are related to responses \(y\) (dependent variables) by some process.

\[
\begin{align*}
Y & \leftarrow \text{Nature} & X
\end{align*}
\]

• A predictive model is some function of the independent variables that mimics this process.

• Predictive Modeling is the process of developing a predictive model.

• There are two goals in fitting a predictive model:
  – Prediction: to be able to predict responses for future input variables \(x\)
  – Information: to understand the process that associates response variables with input variables.

• Note: Just the same as the goal of developing an assumption!!!
What Predictive Modeling Data looks like

Y \rightarrow \text{Nature} \rightarrow X

- The data used reflects this
- Each historical observation is a row in the data
- The dependent response variable $y$ and each independent variable $x$ are columns in the data
- It is common to have a column of weights indicating how much attention should be paid to each observation when fitting the model
  - For example, a policy exposed for 1 year will have twice the influence of a policy exposed for six months, if exposure is chosen as the weight
- A sample taken from the incidence modeling data is included (IncidenceDataSample.xlsx)
  - For this data, (claims/years of exposure) is the response, years of exposure is the weight
- Creating model ready data can be non-trivial
Data exploration is carried out to ensure that the data is appropriate for modeling, i.e.

- All relevant independent variables are present and correctly populated
- The weight is distributed appropriately among the levels of the independent variables
- The response varies appropriately among the levels of the independent variables
- The relationships between the independent variables are reasonable and understood
- The response is appropriately distributed
Preliminary Analyses in Emblem using LTC incidence data

This is done by looking at the following

- Summaries of response and weight by combinations of independent variables
- Analysis of correlations between independent variables
- Analysis of distribution of the response variable
Preliminary Analyses in Emblem using LTC incidence data

Within Emblem,

• Open the Incidence data (File>Open Data File) and press Enter
• Run Correlations Analysis (Modelling>Initial Tests>Correlations)
• What do you notice? Does it make sense?
Preliminary Analyses in Emblem using LTC incidence data

Within Emblem,

- Open Data Analyzer (Modeling>Initial Tests>Data Analyzer)
- Create some Univariate and Bivariate Analyses based on different independent variables
- Do you note anything that surprises you?
In some cases it is important to review the distribution of the response, in order to

- Make sure that it is appropriate, for example
  - Claim costs should be positive, with an appropriate mean, median etc.
  - Proportion of benefits should be between 0 and 1
- Help guide our selection of error structure
  - If response is between 0 and 1, Binomial is best choice
  - If response is negative, Gamma is a bad choice
- This is not useful for the response in the Incidence data
GLMs (Generalized Linear Models) are a flexible, transparent and sophisticated predictive modeling technique.

\[ y = h(\text{Linear Combination of Factors}) + \text{Error} \]

- **Response Variable**
- **Systematic Component**
- **Random Component**

\[ g = h^{-1} \] is called the LINK function and is chosen to reflect the signal most accurately.

Error should reflect underlying process and comes from the exponential family of distributions.

Combination of factors is the model structure.
GLM Building Blocks

\[ y = h(\text{Combination of Factors}) + \text{Error} \]

- **Link function** \((g=h^{-1})\) choice determines how the factors are related to the response in the model
  - **Identity**: Variables related additively (e.g., risk modeling)
  - **Log**: Variables related multiplicatively (e.g., risk modeling)
  - **Logit**: Modeling binary outcomes (e.g., yes/no events)

- Note that if link function is log, then
  - Inverse of Log is Exponential
  - \(\exp(A+B+\ldots+G) = \exp(A) \cdot \exp(B) \cdot \exp(G)\)

Model is multiplicative
GLM Building Blocks

\[ y = h(\text{Combination of Factors}) + \text{Error} \]

- Include variables that are predictive; exclude those that are not
  - Gender may not have major impact on utilization
- Simplify some variables, if full inclusion is not necessary
  - Some levels within a particular predictor may be grouped together
    - e.g. those with incurred age under 50, where incidence is very practically zero
  - A curve may replicate the effect of an ordinal variable e.g. incurred age or duration
- Complicate model structure if the responses depends on combinations of levels of more than one variable (interactions)
  - E.g. the difference between males and females depends on incurred age
GLM Building Blocks

\[ y = h(\text{Combination of Factors}) + \text{Error} \]

- Error functions reflects the variability of the underlying process and can be any distribution within the exponential family of distributions, for example:
  - Gamma consistent with severity modeling
  - Tweedie consistent with pure premium modeling
  - Poisson consistent with incidence modeling
  - Normal useful for a variety of applications

- Appropriate for modeling processes common in insurance
Generally accepted standards are good starting points for link functions and error structures.

<table>
<thead>
<tr>
<th>Observed Response</th>
<th>Most Appropriate Link Function</th>
<th>Most Appropriate Error Structure</th>
</tr>
</thead>
<tbody>
<tr>
<td>--</td>
<td>--</td>
<td>Normal</td>
</tr>
<tr>
<td>Frequency/Mortality/Incidence</td>
<td>Log</td>
<td>Poisson</td>
</tr>
<tr>
<td>Severity/Utilization</td>
<td>Log</td>
<td>Gamma</td>
</tr>
<tr>
<td>Pure Premium</td>
<td>Log</td>
<td>Tweedie</td>
</tr>
<tr>
<td>Retention/Conversion/Termination/</td>
<td>Logit</td>
<td>Binomial</td>
</tr>
<tr>
<td>Cross-sell/Response Rate</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
What we have covered

- **Towers Watson Emblem**
  - What is it and why are we using it for the workshop?
  - Ensuring everyone has access to it

- **What predictive modeling data looks like**

- **Preliminary Analyses in Emblem using LTC incidence data**
  - Correlation analysis
  - Univariate analysis
  - Bivariate analysis
  - Distribution of response (discussion only)

- **Statistical Background to Generalized Linear Models (GLMs)**
Reminder of further sessions

1. Background to the Workshop
2. Predictive Modeling and Applications to Long Term Care Insurance
3. Predictive Modeling of Long Term Care Incidence using Generalized Linear Models in Emblem

We’ll see you on Wednesday, when we will use Emblem to fit a GLM of the incidence data explored today!