Travelers Analytics:
U of M Stats 8053 Insurance Modeling Problem

October 30th, 2013

Nathan Hubbell, FCAS
Shengde Liang, Ph.D.
Agenda

• Travelers: Who Are We & How Do We Use Data?
• Insurance 101
  – Basic business terminology
• Insurance Modeling Problem
  – Introduction
  – Exploratory Data Analysis
  – Assignment Walk-through
How is data used at Travelers?

• Loss, Premium, and Financial Data
• Research & Development
• Unstructured

• Traditional Actuarial Usage
  – Univariate analysis
• Includes external data
  – Multivariate analysis
  – Example: GLMs allow for a non-linear approach in predictive modeling.

• Future development
  – Continued use of sophisticated statistical methods
Insurance 101
Basics of Insurance

Insurance companies sell insurance policies, which are the promise to pay in the event that a customer experiences a loss.

The unique challenge in insurance is that we don’t know what the cost of insuring a customer is when we sell the policy.

Example: The cost to insure an auto customer

It’s impossible to predict if someone is going to

- Get into an accident
- The type of accident (hit a telephone pole, hit another vehicle, bodily injury)
- How bad (cost) the accident will be
To estimate the cost of insuring policyholders, we must predict losses

Two fundamental questions we must answer are:

1. **Ratemaking**: looking to the future
   - Setting rates for policies
   - How much do we need to charge customers for a policy in order to reach our target profit? *Basic idea: price = cost + profit*

2. **Reserving**: looking at the impact of past experience
   - Setting aside reserve money
   - How much money do we need to set aside to pay for claims?

**Note**: We cannot precisely predict losses for each individual or business. However, if we group our customers together, we can build statistical models to predict average loss over a group.
Model Building

- Generalized Linear Models (GLMs)
  - Potential response variables:
    - **Claims – Frequency** (# claims / exposure) (e.g. Poisson, Negative Binomial)
    - **Loss – Severity** (loss $ / claim) (e.g. Gamma, Inverse Gaussian)
    - **Pure Premium** = Frequency * Severity = loss $ / exposure
  - A common **link function** is \( g(x) = \ln(x) \).
  - Probability distribution: **Tweedie**
    - Compound distribution of a **Poisson** claim #
    - And a **Gamma** claim size distribution
    - Large spike at 0 for policies with no claims
    - Wide range of amount in the claims
  - Challenges include:
    - Variable selection
    - Bias-variance trade-off

So what is an example of an actual modeling problem in insurance?

What questions do you have about:

• Travelers?
• Insurance?
• Statistics at Travelers?
Business Problem

• Refer to the one page hand out “Kangaroo Auto Insurance Company Modeling Problem” for more details
• You, as a statistician, work for Kangaroo Insurance, an Australian insurance company
• The underwriter in your company would like you to build a pricing model (pure premium) for the auto insurance product.
• The pricing needs to be competitive.
  – accurately reflect the risk your company is taking.
  – enough segmentation among customers.
• The data from policies written in 2004 and 2005 is provided.
Data Information

- Losses for each vehicle from policies written in 2004 and 2005.
- Each policy was written as one-year originally.
- There are 67,856 policies (vehicles) in the data.
- Ten (10) variables in the data.

<table>
<thead>
<tr>
<th>veh_value</th>
<th>exposure</th>
<th>clm</th>
<th>numclaims</th>
<th>claimcst0</th>
<th>veh_body</th>
<th>veh_age</th>
<th>gender</th>
<th>area</th>
<th>agecat</th>
<th><em>OBSTAT</em></th>
</tr>
</thead>
<tbody>
<tr>
<td>1.06</td>
<td>0.303901</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>HBACK</td>
<td>3</td>
<td>F</td>
<td>C</td>
<td>2</td>
<td>01101</td>
</tr>
<tr>
<td>1.03</td>
<td>0.648871</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>HBACK</td>
<td>2</td>
<td>F</td>
<td>A</td>
<td>4</td>
<td>01101</td>
</tr>
<tr>
<td>3.26</td>
<td>0.569473</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>UTE</td>
<td>2</td>
<td>F</td>
<td>E</td>
<td>2</td>
<td>01101</td>
</tr>
<tr>
<td>4.14</td>
<td>0.317591</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>STNWG</td>
<td>2</td>
<td>F</td>
<td>D</td>
<td>2</td>
<td>01101</td>
</tr>
<tr>
<td>1.38</td>
<td>0.854209</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>HBACK</td>
<td>2</td>
<td>M</td>
<td>A</td>
<td>2</td>
<td>01101</td>
</tr>
<tr>
<td>1.22</td>
<td>0.854209</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>HBACK</td>
<td>3</td>
<td>M</td>
<td>C</td>
<td>4</td>
<td>01101</td>
</tr>
<tr>
<td>1</td>
<td>0.492813</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>HBACK</td>
<td>2</td>
<td>F</td>
<td>C</td>
<td>4</td>
<td>01101</td>
</tr>
<tr>
<td>7.04</td>
<td>0.314853</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>STNWG</td>
<td>1</td>
<td>M</td>
<td>A</td>
<td>5</td>
<td>01101</td>
</tr>
<tr>
<td>1.66</td>
<td>0.4846</td>
<td>1</td>
<td>1</td>
<td>669.51</td>
<td>SEDAN</td>
<td>3</td>
<td>M</td>
<td>B</td>
<td>6</td>
<td>01101</td>
</tr>
<tr>
<td>2.35</td>
<td>0.391513</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>SEDAN</td>
<td>2</td>
<td>M</td>
<td>C</td>
<td>4</td>
<td>01101</td>
</tr>
<tr>
<td>1.51</td>
<td>0.99384</td>
<td>1</td>
<td>1</td>
<td>806.61</td>
<td>SEDAN</td>
<td>3</td>
<td>F</td>
<td>F</td>
<td>4</td>
<td>01101</td>
</tr>
<tr>
<td>0.76</td>
<td>0.539357</td>
<td>1</td>
<td>1</td>
<td>401.8055</td>
<td>HBACK</td>
<td>3</td>
<td>M</td>
<td>C</td>
<td>4</td>
<td>01101</td>
</tr>
<tr>
<td>1.89</td>
<td>0.654346</td>
<td>1</td>
<td>2</td>
<td>1811.71</td>
<td>STNWG</td>
<td>3</td>
<td>M</td>
<td>F</td>
<td>2</td>
<td>01101</td>
</tr>
</tbody>
</table>
• vehicle value, in $10,000s, a numerical variable.
Variable Information – exposure

• The covered period, in years, a numerical variable (always between 0 and 1)
  – The amount of time a vehicle was “exposed” to potential accidents.
Variable Information – clm

• An indicator whether the vehicle/driver had at least one claim during the covered period, 0=No, 1=Yes.
• 4,624/67,856 had at least one claim.
Variable Information – numclaims

- Number of claims during covered period, integer values.
- 4,624/67,856 had at least one claim.

<table>
<thead>
<tr>
<th>Number of Claims</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>63,232</td>
</tr>
<tr>
<td>1</td>
<td>4,333</td>
</tr>
<tr>
<td>2</td>
<td>271</td>
</tr>
<tr>
<td>3</td>
<td>18</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
</tr>
</tbody>
</table>
Variable Information – claimcst0 (target variable)

- The total amount of the claims, in dollars, numeric values.
Variable Information – veh_body

- The vehicle body code, a character string.

CONVT = convertible
HBACK = hatchback
HDTOP = hardtop
MCARA = motorized caravan
MIBUS = minibus
PANVN = panel van
RDSTR = roadster
STNWG = station wagon
UTE - utility
• The age group of insured vehicle, coded as 1, 2, 3, and 4, with 1 being the youngest.
Variable Information – gender

• The gender of driver, F (female) or M (male)

<table>
<thead>
<tr>
<th>Gender</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>F</td>
<td>38,603</td>
</tr>
<tr>
<td>M</td>
<td>29,253</td>
</tr>
</tbody>
</table>
Driver’s area of residence, a character code.

<table>
<thead>
<tr>
<th>Area Code</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>16,312</td>
</tr>
<tr>
<td>B</td>
<td>13,341</td>
</tr>
<tr>
<td>C</td>
<td>20,540</td>
</tr>
<tr>
<td>D</td>
<td>8,173</td>
</tr>
<tr>
<td>E</td>
<td>5,912</td>
</tr>
<tr>
<td>F</td>
<td>3,578</td>
</tr>
</tbody>
</table>
Variable Information – agecat

- Driver’s age category, coded as 1, 2, 3, 4, 5 and 6, with 1 being the youngest.

<table>
<thead>
<tr>
<th>Driver Age Category</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5,742</td>
</tr>
<tr>
<td>2</td>
<td>12,875</td>
</tr>
<tr>
<td>3</td>
<td>15,767</td>
</tr>
<tr>
<td>4</td>
<td>16,189</td>
</tr>
<tr>
<td>5</td>
<td>10,736</td>
</tr>
<tr>
<td>6</td>
<td>6,547</td>
</tr>
</tbody>
</table>
Questions May Be Asked

• What models did you fit?
  – what is your assumption(s)?
  – is your assumption reasonable?
  – how do you check your assumption(s)?
• What is the impact of each variable?
  – are all variables equally important?
  – if not, which ones are more important? How do you measure it?
• How do you check your model actually works (generalizability)?

What questions do you have about the “Kangaroo Insurance Company Modeling Problem”?
References and Resources

• Contacts
  – Nathan Hubbell – N H U B B E L L @ t r a v e l e r s . c o m
  – Shengde Liang – S L I A N G @ t r a v e l e r s . c o m

• Travelers Careers
  – http://www.travelers.com/careers
  – Actuarial and Analytics Research Internship and Full Time

• A Practitioner’s Guide to Generalized Linear Models