GLM I:

An Introduction to Generalized Linear Models

Prepared For: Casualty Actuarial Society
2009 Ratemaking & Product Management Seminar

Prepared By: Paul D. Anderson, FCAS, MAAA
paul.anderson@milliman.com
(262) 641-3531

March 10, 2009

Antitrust Notice

The Casualty Actuarial Society is committed to adhering strictly to the letter and spirit of the antitrust laws. Seminars conducted under the auspices of the CAS are designed solely to provide a forum for the expression of various points of view on topics described in the programs or agendas for such meetings.

Under no circumstances shall CAS seminars be used as a means for competing companies or firms to reach any understanding – expressed or implied – that restricts competition or in any way impairs the ability of members to exercise independent business judgment regarding matters affecting competition.

It is the responsibility of all seminar participants to be aware of antitrust regulations, to prevent any written or verbal discussions that appear to violate these laws, and to adhere in every respect to the CAS antitrust compliance policy.

Overview of Presentation

- ØPreparing the Data for Analysis
- ØPreliminary Analysis of the Data
- ØSampling the Data
- ØRunning a GLM

Preparing the Data for Analysis

How Much Data is Needed?

- In summary … as much as possible
  - Ø Large datasets can always be sampled or made smaller
  - Ø If more data is needed, consider pulling more years
    - Assumes additional years are readily available
    - Assumes older years contain valid information in data fields
- Preferred threshold is a minimum of 5,000 claims
  - Ø Depends on current level of sophistication & volatility of data
    - Analysis Y began with 120,000 records & 4,000 claims
    - Still produced meaningful results on a volatile line of business

Placing Data in Proper Platform

- Format of the Data being Retrieved / Received
  - Ø Text (.txt) or Flat (.dat) Files
    - Fixed length – each record is a pre-defined length
    - Delimited (tab, comma, etc.) – each variable is separated by a common character
  - Ø Other Database Files: Excel, Access, SQL, etc.
- Statistical Software / Platform
  - Ø SAS is widely used & easy to understand
  - Ø Others include S, S+, and R

Understanding the Data

- Initial Data Check
  - Ø Ask for control totals & documentation of variables
- How was the Data Compiled?
  - Ø Raw data dump vs. Combination of multiple files
- Secondary Data Check
  - Ø Match totals in your dataset to control totals
  - Ø Did you receive a complete dataset?
  - Ø Frequency distributions of key fields
    - Years, States, Companies, etc.

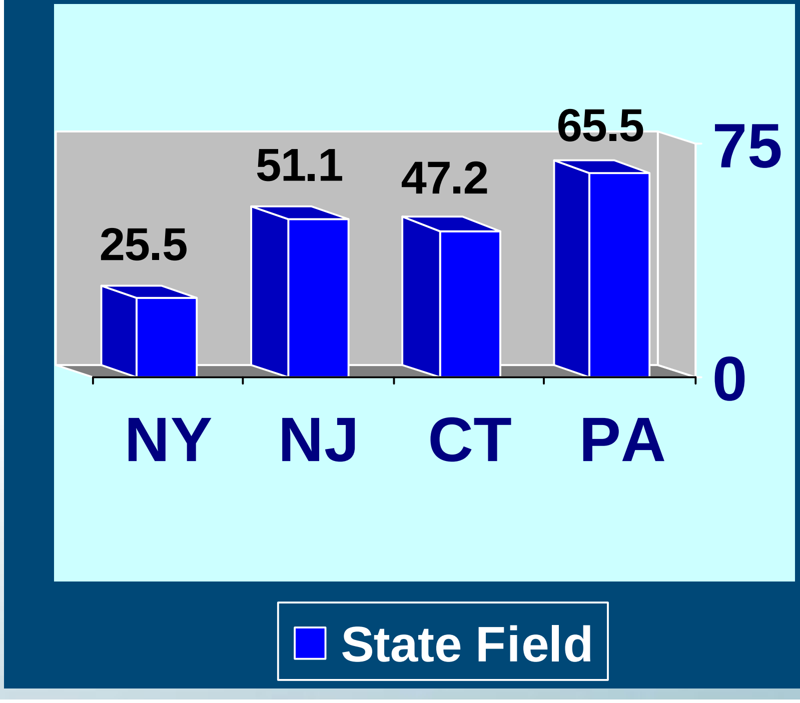
Understanding the Data (cont.)

- Review Data Fields Provided
  - Ø Did you receive all the needed fields?
  - Ø Expected values within each field
  - Ø Completeness of data within each field
    - Are any data fields missing for 100% of records?
  - Ø Unique Identifiers
  - Ø Linking Variables
  - Ø Defined Variables vs. Derived Variables

Understanding the Data (cont.)

Distribution of State Field

WC Claims in Thousands



Distribution of Variable X

(Number of Records)

	Freq	%	Cum. Freq	Cum. %
1	14,251	8.2	14,251	8.2
2	11,251	6.5	25,502	14.7
3	11,549	6.6	37,051	21.3
4	55,642	32.0	92,693	53.3
5	44,654	25.7	137,347	78.9
6	22,111	12.7	159,458	92.6
AA	14,541	8.4	173,999	100.0

Missing = 25,000

Examples of Data Problems

- Problem 1: Duplicate Records
  - Ø Definition of Problem:
    - Multiple records for the same policy, usually with different information applying to each record
  - Ø How to Identify Problem:
    - First “dot”/Last “dot”
  - Ø Solution:
    - Nodupkey
    - Last “dot”

Examples of Data Problems (cont.)

- Problem 2: Missing Records (Incomplete Data)
  - Ø Definition of Problem:
    - Missing some or all of data expected
  - Ø How to Identify Problem:
    - Frequency distributions and control totals
  - Ø Solution:
    - Request data resubmission from data source

Examples of Data Problems (cont.)

- Problem 3: Fields with Missing Values
  - Ø Definition of Problem:
    - A field has missing values for all or most records
  - Ø How to Identify Problem:
    - Frequency distributions and/or univariate analysis
  - Ø Solution:
    - Confirm whether field is rarely used or if this indicates a larger data issue
    - May require resubmission from data source
    - Eliminate field from dataset as it will not be useful in the modeling process

Other Possible Data Problems

- Unexpected values reported for a field
- Error in data compilation (usually when combining multiple data sources)
- Extraneous data provided
  - Not a problem, just reminder to check for and eliminate unnecessary data as early in the process as possible.

Preliminary Analysis of the Data

Identification of Key Components

- Define Modeling Variable (i.e. Loss Ratio)
- Develop the list of “Contenders”
  - Ø What fields might you want to model?
- Compile Base Variables (Premium, Losses)

Univariate Analysis:

A Preliminary Step in Analyzing the Data

- Continues process of reducing the list of contenders
  - Ø File size is dependent on number of variables
  - Ø The sooner you can eliminate variables from the analysis, the more manageable your data becomes.
- Further checks fields and values for data issues
  - Ø i.e. low loss ratio might indicate incorrect compilation of losses and/or premiums
- Identifies potential groupings within variables

Sample Univariate Analysis

Flag 1	# of Policies	Premium (Millions)	Losses (Millions)	Avg. Prem	Prem Rel.	Loss Ratio	LR Rel.
Y	355,585	$210	$99	$591	0.92	0.471	1.17
N	55,546	$55	$8	$990	1.54	0.145	0.36

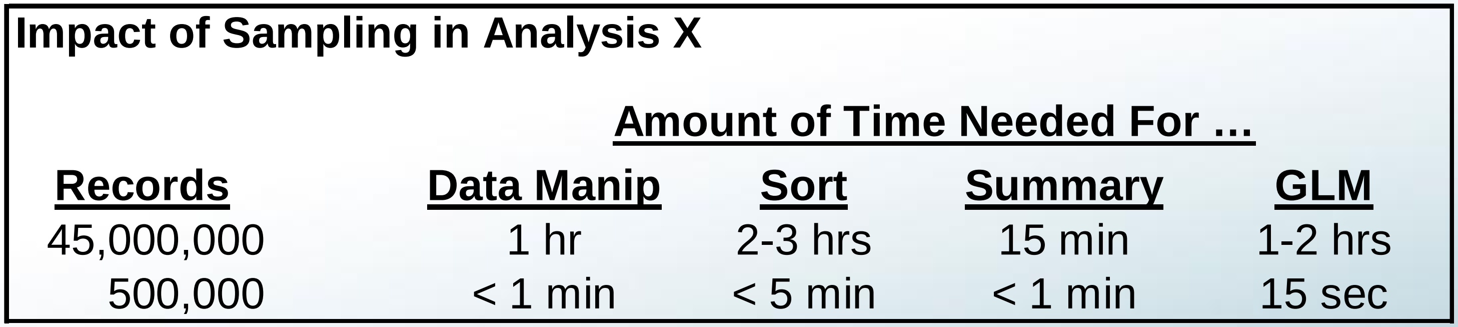
What we observe:

- Since Flag 1 = “N”, with a loss ratio relativity of 0.36, differs significantly from Flag 1 = “Y” and from an average relativity of 1.00, this field would be considered as a potential contender in the model

Sampling the Data

Reasons to Sample the Data

1. Reduce the size of the overall database
   - Large databases are preferred for predictive modeling, but it is possible to be too big
   - Improves efficiency of programs and productivity of analyst



Impact of Sampling in Analysis X

	Amount of Time Needed For ...			
Records	**Data Manip**	**Sort**	**Summary**	**GLM**
45,000,000	1 hr	2-3 hrs	15 min	1-2 hrs
500,000	< 1 min	< 5 min	< 1 min	15 sec

2. Validate the model being built
   - Otherwise, modeling process simply explains history and may not be the best predictor of the future

Sampling Methods in Predictive Modeling

- Random Sampling
  - Ø Assign a random number to each record and divide these random numbers into groups
  - Ø Each record has an equal probability of being selected
  - Ø Goal is to represent the population
- Systematic Sampling
  - Ø Selecting a subset of the data using specified criteria
    - Every 10th record
    - Policy numbers ending in “X”
  - Ø Easy to implement & efficient, but assumes database is already random

Sampling Methods (cont.)

- Sampling for Purposes of Validation
  1. Sampling
     - Sample created & set aside (20-40% of total data)
     - Model built on remaining data (variables selected & preliminary parameter estimates)
     - Validate that model works on sample or use sample to choose between several alternative models
     - Finalize model using all data (final parameter estimates)

Sampling Methods (cont.)

- Sampling for Purposes of Validation (cont.)

  2. Resampling
     - Used for smaller databases
     - Similar to Sampling, but repeated N times using N different samples
     - Variables are selected that are most robust and remain predictive across all (or most) of the N iterations
     - After N iterations of modeling, finalize model using all data
  3. Partitioning
     - Alternative method for smaller databases
     - Similar to Resampling, but data divided into N partitions
     - 1st partition set aside, model built & validated against 1st partition
     - Process repeated for other N-1 partitions
     - Finalize model using all data

Running a GLM

Data Needed for a GLM

In summary … as much detail as possible

- Ø Level of detail depends on the goal of the analysis
- Ø One record per policy, per year
  - Useful for Pricing, Ratemaking, or Underwriting analyses
- Ø One record per claim, possibly per evaluation period
  - Useful for Claim-related analyses
- Ø One record per agent, per year
  - Useful for Sales-related analyses
- Ø Each record should contain data related to the dependent variable being modeled & as many independent variables as possible

What Can be Modeled With a GLM?

Anything that we try to predict or estimate:

- Ø Pure Premiums
- Ø Loss Ratios
  - Can be complex due to historical changes in class plan
- Ø Claim Frequencies
- Ø Claim Severities
  - Unlimited or capped
- Ø Retention Ratios / Termination Ratios
- Ø Close Ratios
- Ø Claim Settlement Patterns
- Ø Relativities of any of the above

Inputs & Outputs of a GLM

Inputs:

- Ø Database at the appropriate level of detail
  - Data/Records have been “cleaned up”, filtered, tested, & sampled

Outputs:

- Ø Listing of values for each variable being modeled
  - Parameter estimate
  - 3-4 statistical measures to help identify confidence in each estimate
  - Note: Last value within each variable is usually the base class (i.e. factor = 1.00)

Selecting a Final Model

- Consider reduction in residual (error) vs. added complexity of an another variable
- Balance between predictive and explanatory
    - Ø Overall mean is predictive
    - Ø Individuality is explanatory

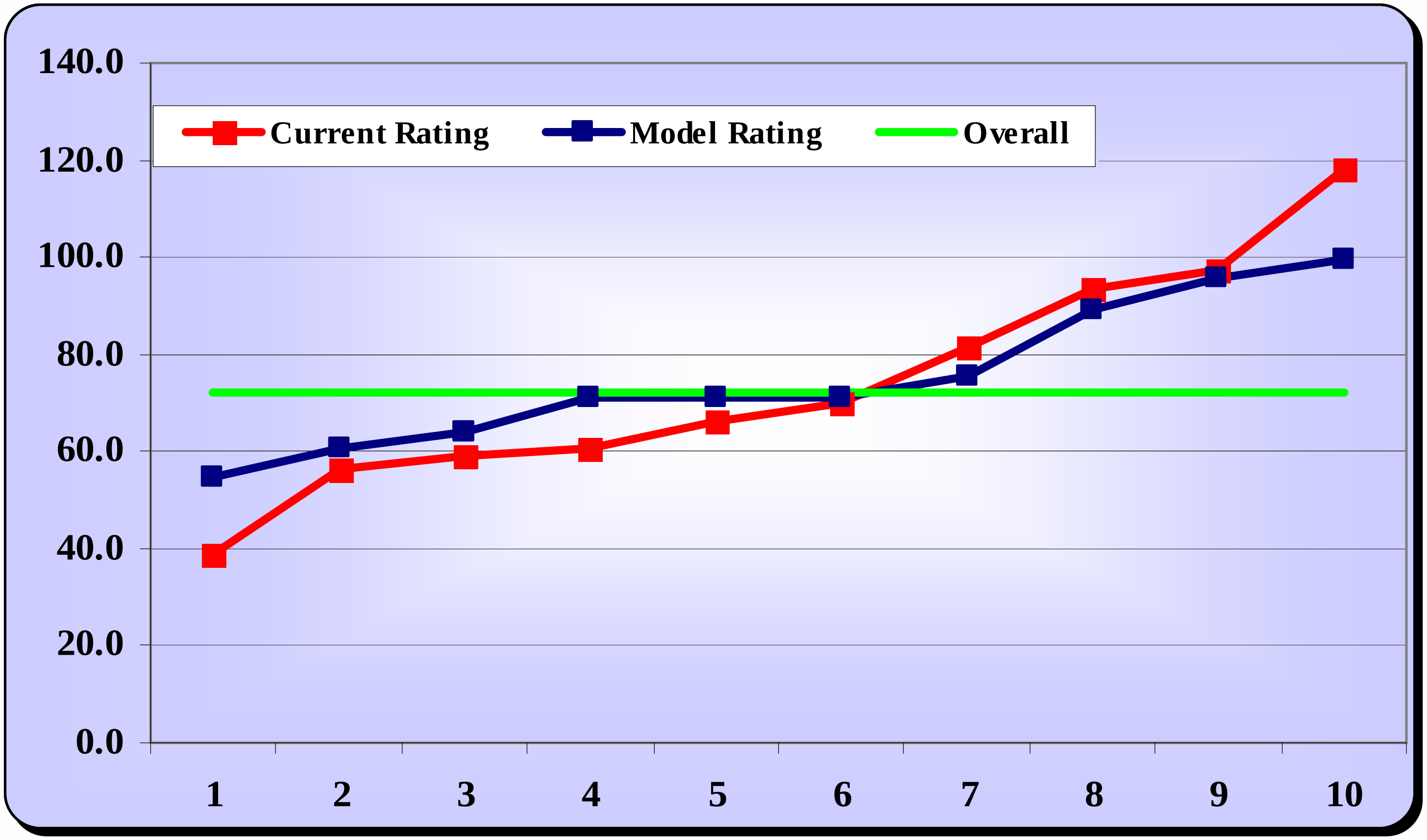
Sample table showing process of monitoring results:

	Run 1	Run 2	Run 3	Run 4	Run 5
Log Likelihood	-3,041	-3,025	-3,017	-3,010	-3,002
Degrees of Freedom	4	7	9	14	25
-2 x (Chng in Log Likelihood)		30	17	13	16
ChiSq = Pr(Improvement NOT Signif)		0.00%	0.02%	2.11%	15.08%

Validating the Results

- Goal of validation is to ensure that parameter estimates in selected model truly are good predictors

Sample loss ratio chart shows reduced subsidy by decile


Current Rating
Model Rating
Overall