GLM II: Basic Modeling Strategy

CAS Predictive Modeling Special Interest Seminar

Geoff Werner – Senior Consultant – EMB America

EMB

CAS Predictive Modeling

Oct 2006

EMB

Basic Modeling Session

PURPOSE: To discuss basic modeling strategies and techniques for building appropriate GLM models

OUTLINE

- Background
- Overall Modeling Strategy
- Basic Predictive Modeling Steps
  1. Get clean data
  2. Select an initial error structure, link function, and model structure
  3. Test error structure/link function
  4. Preliminary investigation
  5. Build predictive models iteratively
  6. Validate final predictive model
  7. Combine models, if modeling frequency and severity
- Summary

CAS Predictive Modeling

Oct 2006

- **Background**
- Overall Strategy
- Modeling Steps
  1. Get Data
  2. Initial Sels
  3. Test Error/Link
  4. Preliminary Investigation
  5. Build Models
  6. Validate Models
  7. Combine Models
- Summary

Purpose of Predictive Modeling

- To predict a response variable using a series of explanatory variables (or rating factors)

Independent/Predictors		Weights	Dependent/Response
Age	Accidents	Claims	Losses
Limit	Convictions	Exposures	Claims
Territory	Credit Score	Premium	Retention

↓

Statistical Model

↓

Model Results
Parameters
Validation Statistics

Same techniques apply regardless of what is being modeled, this session will focus on risk modeling as it is the most common application

EMB

CAS Predictive Modeling

Oct 2006

- **Background**
- Overall Strategy
- Modeling Steps
  1. Get Data
  2. Initial Sels
  3. Test Error/Link
  4. Preliminary Investigation
  5. Build Models
  6. Validate Models
  7. Combine Models
- Summary

Generalized Linear Models (GLMs)

- Multivariate method that considers all factors simultaneously

Response Variable = Systematic Component + Random Component

y = h(Linear Combination of Rating Factors) + Error

$g = h^{-1}$ is called the LINK function

Error reflects underlying process

Combination of rating factors is the model structure

EMB

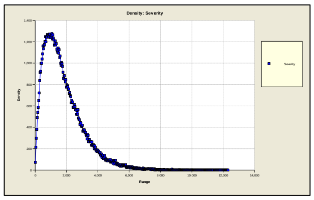
CAS Predictive Modeling
Oct 2006

GLM Building Blocks
Error Structure

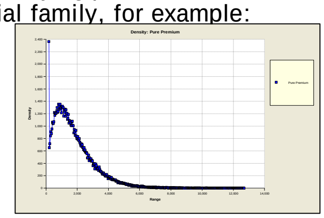
y = h(Linear Combination of Rating Factors) + **Error**

- Background
- Overall Strategy
- Modeling Steps
  1. Get Data
  2. Initial Sets
  3. Test Error/Link
  4. Preliminary Investigation
  5. Build Models
  6. Validate Models
  7. Combine Models
- Summary

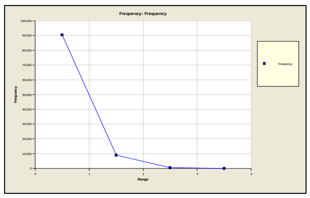
- Reflects the variability of the underlying process and can be any distribution within the exponential family, for example:



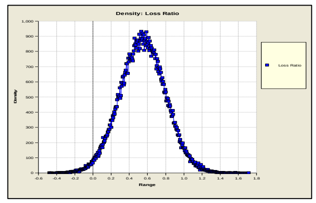
- Gamma consistent with severity modeling, may want to try Inverse Gaussian



- Tweedie consistent with pure premium modeling



- Poisson consistent with frequency modeling



- Normal useful for a variety of applications

EMB

CAS Predictive Modeling
Oct 2006

GLM Building Blocks
Model Structure

y = h(**Linear Combination of Rating Factors**) + Error

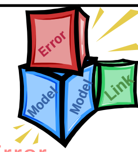
- Background
- Overall Strategy
- Modeling Steps
  1. Get Data
  2. Initial Sets
  3. Test Error/Link
  4. Preliminary Investigation
  5. Build Models
  6. Validate Models
  7. Combine Models
- Summary

- Include variables that are predictive, exclude those that are not
  - Gender may not have major impact on theft severity
- Simplify some rating factors, if full inclusion is not necessary
  - Some levels within a particular predictor may be grouped together (e.g., 50-54 year olds)
  - A curve may replicate the signal (e.g., amount of insurance)
- Complicate model if the relationship between levels of one variable depends on another characteristic
  - The difference between males and females depends on age

EMB

GLM Building Blocks

Link Function

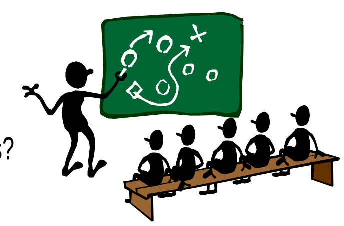


y = **h(**Linear Combination of Rating Factors**)** + Error

- Link function (g=h-1) chosen to based on how the factors are related to produce the best signal:
  - Log: variables related multiplicatively (e.g., risk modeling)
  - Identity: variables related additively (e.g., risk modeling)
  - Logit: retention or risk modeling
  - Reciprocal: canonical link for gamma distribution
  - Mixed: additive/multiplicative rating algorithms

EMB

Overall Modeling Strategy Questions



- Should you model loss ratios?
- Should you model frequency and severity separately by coverage/peril or model in the aggregate?
- Should you only model current rating variables?

EMB

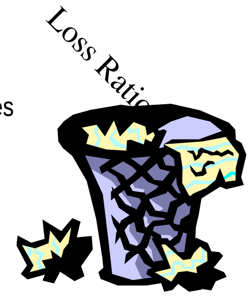
CAS Predictive Modeling

Oct 2006

Should You Model Loss Ratios?

- Background
- **Overall Strategy**
- Modeling Steps
  1. Get Data
  2. Initial Sels
  3. Test Error/Link
  4. Preliminary Investigation
  5. Build Models
  6. Validate Models
  7. Combine Models
- Summary

EMB



- Some companies model loss ratios
  - May find it difficult to obtain exposures
  - Do not want to pull all of the data, so assume using loss ratios will "adjust" for excluded variables
  - Habit formed when performing traditional analysis
- Theoretical and practical *disadvantages* to loss ratio modeling
  - On-level calculations
  - No defined error distribution
  - Difficult to distinguish noise from pattern
  - If changes made, models cannot be reused

CAS Predictive Modeling

Oct 2006

Loss Ratio Modeling

On-Level Calculations

- Background
- **Overall Strategy**
- Modeling Steps
  1. Get Data
  2. Initial Sels
  3. Test Error/Link
  4. Preliminary Investigation
  5. Build Models
  6. Validate Models
  7. Combine Models
- Summary

EMB

- When modeling using loss ratios, premiums should be put on-level to adjust for changes during or after the historical period
  - Rate changes
  - Underwriting changes
- Not sufficient to use an average on-level approach (e.g., parallelogram method) when changes impact classes differently
- Instead, put premiums on-level at the granular level (e.g., extension of exposures)
  - Can be time consuming
  - Data may not be readily available
- Depending on the type and magnitude of the changes, failure to put premiums on level can result in serious under- and over-predictions
- Pure premiums use exposures so this is a non-issue

CAS Predictive Modeling

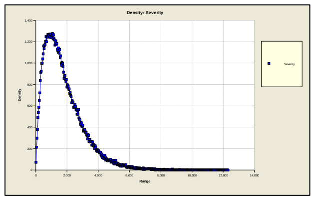
Oct 2006

Loss Ratio Modeling

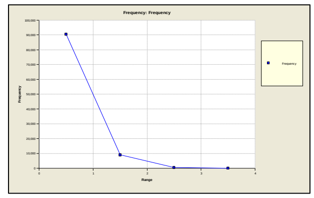
Defined Error Structure

- Background
- **Overall Strategy**
- Modeling Steps
  1. Get Data
  2. Initial Sels
  3. Test Error/Link
  4. Preliminary Investigation
  5. Build Models
  6. Validate Models
  7. Combine Models
- Summary

- When modeling frequency and severity, there are generally accepted distributions



Gamma considered a standard for severity modeling



Poisson considered a standard for frequency modeling

- What is the typical distribution for loss ratios?
  - There is no generally accepted standard
  - The distribution will vary by company, line, and over time

EMB

CAS Predictive Modeling

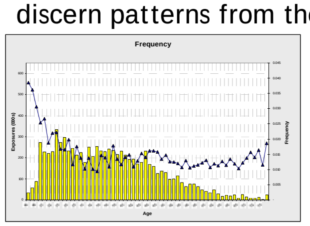
Oct 2006

Loss Ratio Modeling

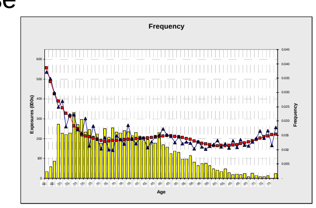
Discerning Patterns

- Background
- **Overall Strategy**
- Modeling Steps
  1. Get Data
  2. Initial Sels
  3. Test Error/Link
  4. Preliminary Investigation
  5. Build Models
  6. Validate Models
  7. Combine Models
- Summary

- When viewing frequency and severity data separately, easy to discern patterns from the noise

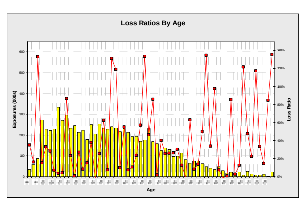


Raw Frequency by Age of Driver



Smoothed Frequency by Age of Driver

- With loss ratio difficult or impossible to determine pattern from noise



Raw Loss Ratio by Age of Driver

EMB

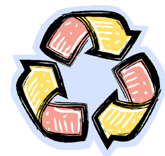
CAS Predictive Modeling

Oct 2006

- Background
- **Overall Strategy**
- Modeling Steps
  1. Get Data
  2. Initial Sels
  3. Test Error/Link
  4. Preliminary Investigation
  5. Build Models
  6. Validate Models
  7. Combine Models
- Summary

Loss Ratio Modeling

Re-usability



- Loss ratio modeling
  - Modeling losses/premiums, thus it is imperative that premiums be put on-level
  - If a review results in changes
    - All of the loss ratios will change
    - The relationships between levels of factors may change as well
  - Models built in last review will be inappropriate
- Pure Premium modeling
  - Modeling does not involve premium, thus unnecessary to put premiums on level
  - If a review results in changes
    - The frequencies, severities, pure premiums will not change
    - The relationships between levels will be unaffected
  - Models built in last review may still be appropriate

CAS Predictive Modeling

Oct 2006

- Background
- **Overall Strategy**
- Modeling Steps
  1. Get Data
  2. Initial Sels
  3. Test Error/Link
  4. Preliminary Investigation
  5. Build Models
  6. Validate Models
  7. Combine Models
- Summary

Granular or Combined Modeling?

- Some tempted to model raw pure premiums or combined coverages/perils, presumably to save time
- As with traditional analysis (e.g., selecting loss trends), preferable to analyze at the granular level

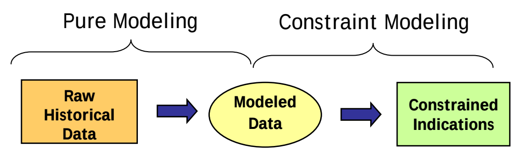
Freq/Severity or Pure Premium	By-Peril or All Perils
Severity trends mask frequency signal	High variable perils mask stable perils
Predictors impact frequency and severity differently (e.g., limit)	Predictors affect perils differently (e.g., theft device)
Frequency and severity have defined error structures	Perils have different size of loss distributions
Different frequency and severity trends can mask results	Different loss trends by peril can mask results

- If necessary, consider Tweedie and Joint Modeling macros

- Background
- **Overall Strategy**
- Modeling Steps
  1. Get Data
  2. Initial Sels
  3. Test Error/Link
  4. Preliminary Investigation
  5. Build Models
  6. Validate Models
  7. Combine Models
- Summary

Use All Available Data?

- Companies may limit number of variables reviewed. For example, companies may mistakenly exclude
  - Variables not allowed by regulation or not currently used
  - Variables not being changed with current review
  - Underwriting variables



<u>Pure Modeling</u>

- Use all data to remove "noise" and find signal
- Example, geodemographic data may be more predictive than current territory

<u>Constraint Modeling</u>

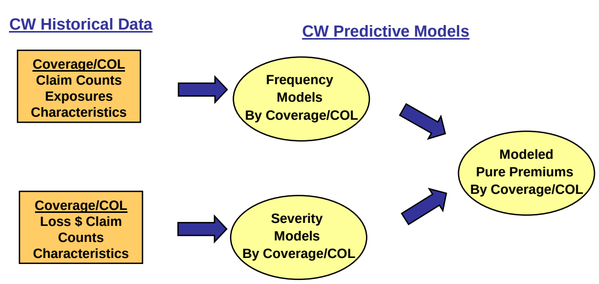
- Convert modeled results into usable indications
- Incorporate restrictions
  - Systems
  - Regulatory
  - Competitive

EMB

CAS Predictive Modeling
Oct 2006

- Background
- **Overall Strategy**
- Modeling Steps
  1. Get Data
  2. Initial Sels
  3. Test Error/Link
  4. Preliminary Investigation
  5. Build Models
  6. Validate Models
  7. Combine Models
- Summary

Predictive Modeling Overall Strategy



- Avoid modeling loss ratios
- Build frequency and severity models by coverage/cause of loss
- Use all available data to find the best signal

EMB

CAS Predictive Modeling

Oct 2006

Basic Modeling Steps

1 Get Clean Data → 2 Make Initial Selections → 3 Test Error Structure & Link Function → 4 Prelim Invest → 5 Build Model Structure → 6 Validate Models → 7 Combine Models

- Background
- Overall Strategy
- **Modeling Steps**
  1. Get Data
  2. Initial Sels
  3. Test Error/Link
  4. Preliminary Investigation
  5. Build Models
  6. Validate Models
  7. Combine Models
- Summary

1. Gather necessary internal and external data
2. Select initial error structure, link function, and model structure
3. Perform basic diagnostic tests to become familiar with data
4. Validate initial selections for error structure and link function
5. Build predictive models
   - Add/exclude variables
   - Group levels
   - Include variates
   - Add interactions
6. Perform tests to validate the models built
7. Combine granular models, if necessary

EMB

CAS Predictive Modeling

Oct 2006

Get Clean Data

- Background
- Overall Strategy
- **Modeling Steps**
  1. **Get Data**
  2. Initial Sels
  3. Test Error/Link
  4. Preliminary Investigation
  5. Build Models
  6. Validate Models
  7. Combine Models
- Summary

- Good project results start with good data
  - Internal data
  - External data
- Data remains the number 1 issue
  - Null records or bad data, especially for variables not used in rating
  - Poor linkage between losses and policy characteristics
  - Too much pre-banding of data
  - No mapping of old groupings into new groupings
  - For auto, no linkage between operator, vehicle, and policy characteristics
  - Inconsistency between variables (e.g., 30 year olds living in a retirement community)
- Key: spend the right amount of time on data acquisition!
  - Typically 50% of first review
  - Some issues cannot be resolved, impact on analysis depends on the type and extent of the problem

EMB

Initial Selections

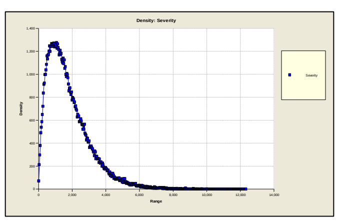
- Use generally accepted standards as starting point for link functions and error structures

Observed Response	Most Appropriate Link Function	Most Appropriate Error Structure	Variance Function
--	--	Normal	μ^0
Claim Frequency	Log	Poisson	μ
Claim Severity	Log	Gamma	μ^2
Claim Severity	Log	Inverse Gaussian	μ^3
Risk Premium	Log	Gamma or Tweedie	μ^T
Retention Rate	Logit	Binomial	$\mu (1-\mu)$
Conversion Rate	Logit	Binomial	$\mu(1- \mu)$

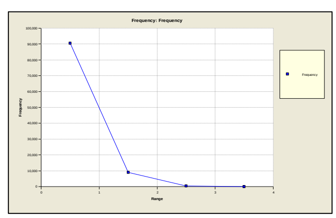
- Reasonable starting point for model structure
  - All or all known important factors
  - Prior model (last year or other related peril)
  - Forward regression model

Test Error Structure/Link Function
Distribution Analysis

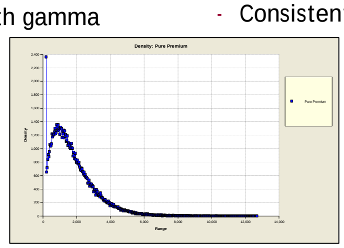
- Examine plots of the data (e.g., size of loss distribution)



- Consistent with gamma



- Consistent with Poisson



- Consistent with Tweedie

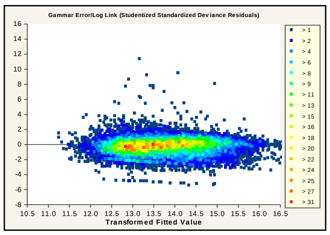
CAS
Predictive
Modeling
Oct 2006

Test Error Structure/Link Function

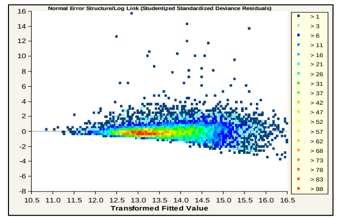
Macro Residual Analysis

EMB

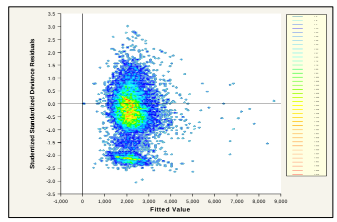
- Plot of all residuals tests selected error structure/link function



- Elliptical pattern is ideal



- Fanning out suggests power of variance function is too low



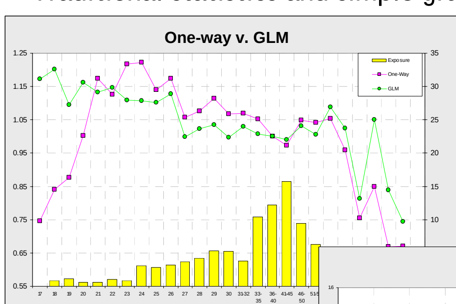
- Two concentrations suggests two perils: split of use joint modeling

CAS
Predictive
Modeling
Oct 2006

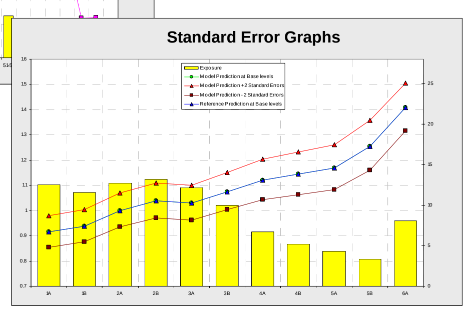
Preliminary Investigation

EMB

- Traditional statistics and simple graphs provide “quick” feel



- Highlights what others within your company “know”

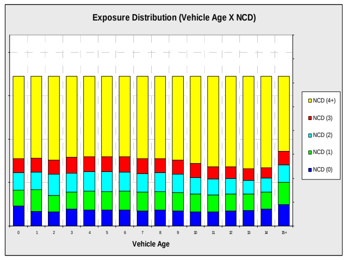


- Quickly highlight trends in your data

Preliminary Investigation

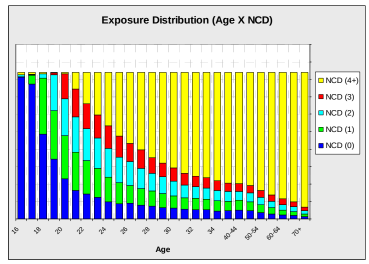
- Background
- Overall Strategy
- **Modeling Steps**
  1. Get Data
  2. Initial Sels
  3. Test Error/Link
  4. **Preliminary Investigation**
  5. Build Models
  6. Validate Models
  7. Combine Models
- Summary

- Statistics can (e.g., Cramer's V) identify correlated variables



Low Correlation (.025)

- Distribution of number of years claims free about the same for each vehicle age



High Correlation (.253)

- Older drivers are more likely to be claim-free

- Identifies independent variables that will have an effect on each other

EMB

CAS Predictive Modeling
Oct 2006

Building the "Best" Model

- Background
- Overall Strategy
- **Modeling Steps**
  1. Get Data
  2. Initial Sels
  3. Test Error/Link
  4. Preliminary Investigation
  5. **Build Models**
  6. Validate Models
  7. Combine Models
- Summary

- To produce a sensible model that explains recent historical experience and is likely to be predictive of future experience

Overall Mean ↓

"Best" Models

1 parameter for each observation ↓

Model Complexity
(Number of Parameters)

↑ **UNDERFIT**

Predictive

Poor explanatory power

↑ **OVERFIT**

Poor predictive power

Explains history

EMB

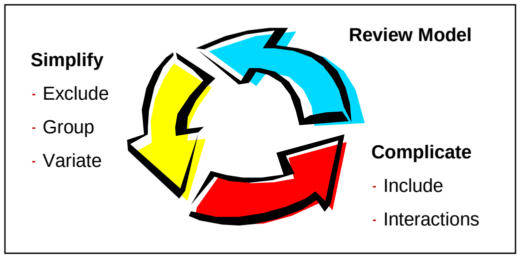
CAS Predictive Modeling
Oct 2006

- Background
- Overall Strategy
- **Modeling Steps**
  1. Get Data
  2. Initial Sels
  3. Test Error/Link
  4. Preliminary Investigation
  5. **Build Models**
  6. Validate Models
  7. Combine Models
- Summary

EMB

Building the "Best" Model

- Modeling is an iterative process



- How does the analyst decide the "Best" Model?
  - Parameters/standard errors
  - Consistency of patterns over time or random data sets
  - Type III statistical tests (e.g., X^2 tests)
  - Judgment (e.g., do the trends make sense)

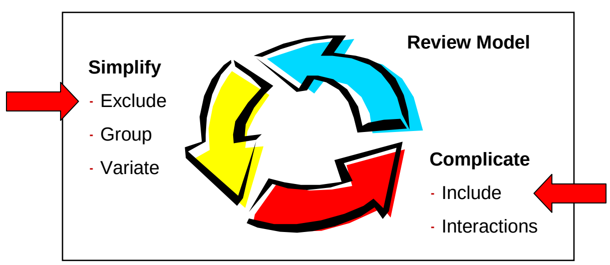
CAS Predictive Modeling
Oct 2006

- Background
- Overall Strategy
- **Modeling Steps**
  1. Get Data
  2. Initial Sels
  3. Test Error/Link
  4. Preliminary Investigation
  5. **Build Models**
  6. Validate Models
  7. Combine Models
- Summary

EMB

Building the "Best" Model

- Modeling is an iterative process



- Add/Exclude: does the independent variable have predictive power that warrants including it in the model?

CAS Predictive Modeling

Oct 2006

- Background
- Overall Strategy
- **Modeling Steps**
  1. Get Data
  2. Initial Sels
  3. Test Error/Link
  4. Preliminary Investigation
  5. **Build Models**
  6. Validate Models
  7. Combine Models
- Summary

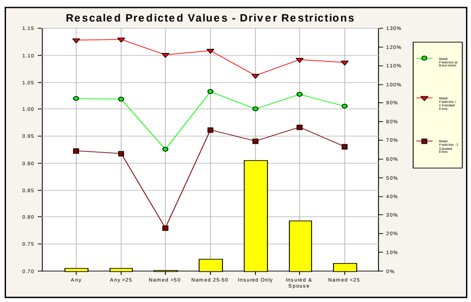
EMB

Build Models

Include/Exclude Factors

- Parameters/standard errors tell importance of factors and "confidence" in estimates
  - If all the parameters are essentially the same or have large standard errors, the factor may not be important

Name	Value	Standard Error	Standard Error (%)	Exp(Value)
Any	0.0174	0.04183	240.8	1.0175
Any>25	0.0212	0.04349	205.4	1.0214
Named >50	-0.0961	0.08120	84.5	0.9084
Named 25-50	0.0357	0.02194	61.4	1.0364
Insured Only				
Insured & Spouse	0.0255	0.01272	49.8	1.0259
Named <25	-0.0446	0.02663	59.7	0.9564



- Graph of parameters/standard errors and "horizontal line test" identifies importance of a factor

CAS Predictive Modeling

Oct 2006

- Background
- Overall Strategy
- **Modeling Steps**
  1. Get Data
  2. Initial Sels
  3. Test Error/Link
  4. Preliminary Investigation
  5. **Build Models**
  6. Validate Models
  7. Combine Models
- Summary

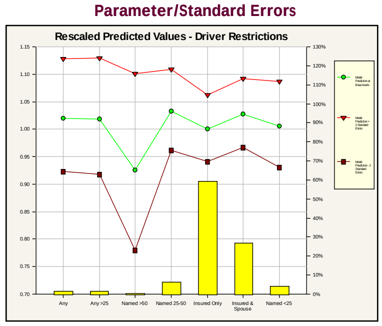
EMB

Build Models

Include/Exclude Factors

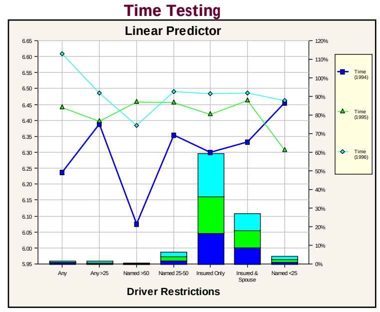
- Examine consistency over time or random parts of data

Parameter/Standard Errors



- Main effects graph may indicate a questionable estimate

Time Testing



- By testing the pattern over time can see if the same thing happens each year

CAS Predictive Modeling Oct 2006

Build Models
Include/Exclude Factors

- Goodness of fit tests (e.g., Chi-Squared) can be used to determine the explanatory power of a variable
  - Null hypothesis is that the models with and without the factor are the same

Chi-Squared

Model	With	Without
Deviance	8,906.4414	8,909.6226
Degrees of Freedom	18,469	18,475
Scale Parameter	0.4822	0.4823
Chi Square Test		78.6%

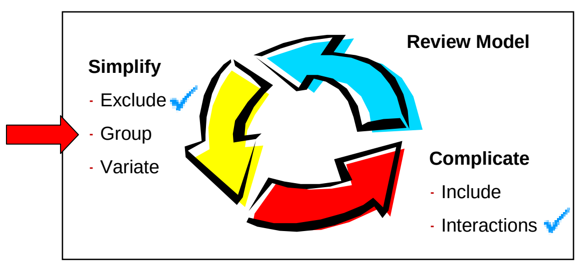
Score	H_0	Indicated Model
<5%	Reject	More Complex: Include Factor
5%-30%	???	???
>30%	Accept	Simpler: Exclude Factor

EMB

CAS Predictive Modeling Oct 2006

Building the “Best” Model

- Modeling is an iterative process



- Group: should some of the levels of a given variable be combined?

EMB

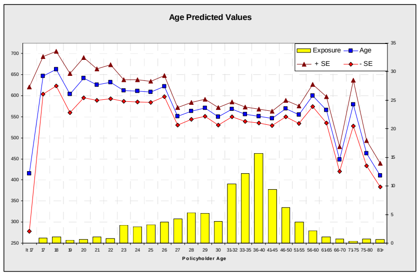
CAS Predictive Modeling
Oct 2006

- Background
- Overall Strategy
- **Modeling Steps**
  1. Get Data
  2. Initial Sets
  3. Test Error/Link
  4. Preliminary Investigation
  5. **Build Models**
  6. Validate Models
  7. Combine Models
- Summary

EMB

Build Models
Group Rating Levels

- Parameters/standard errors tell importance of varying estimates for each level



- Similar parameters or "plateaus" indicate potential groups
- Look for low weights

Name	Value	Standard Error	Standard Error (%)	Weight	E(Value)
Lt 17	-0.2872	0.40047	139.4	3	0.7504
17	0.1597	0.06488	40.6	162	1.1731
18	0.1838	0.05642	30.7	211	1.2018
19	0.0915	0.07222	78.9	106	1.0958
20	0.1506	0.07009	46.6	111	1.1625
21	0.1254	0.05478	43.7	195	1.1336
22	0.1364	0.05916	43.4	156	1.1462
23	0.1038	0.03476	33.5	587	1.1094
24	0.1022	0.03559	34.8	539	1.1076
25	0.0979	0.03288	33.6	602	1.1029
26	0.1207	0.03098	25.7	700	1.1283
27	-0.0015	0.02947	1,929.7	795	0.9985
28	0.0221	0.02635	119.0	1,004	1.0224
29	0.0345	0.02611	75.7	983	1.0351
30	-0.0021	0.02925	1,396.1	711	0.9979
31-32	0.0291	0.02059	70.8	1,952	1.0295
33-35	0.0079	0.01941	244.6	2,294	1.0080
36-40				2,953	
41-45	-0.0103	0.02110	204.5	1,769	0.9897

- Group levels with
  - Base level
  - Neighboring classes

CAS Predictive Modeling
Oct 2006

- Background
- Overall Strategy
- **Modeling Steps**
  1. Get Data
  2. Initial Sets
  3. Test Error/Link
  4. Preliminary Investigation
  5. **Build Models**
  6. Validate Models
  7. Combine Models
- Summary

EMB

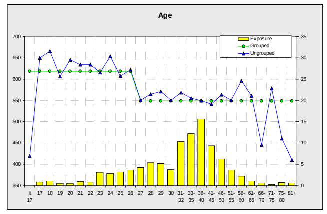
Build Models
Group Rating Levels

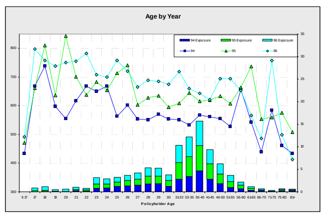
- Standard errors discussed earlier identify levels that should be grouped with the base class
- Standard error of the parameter differences identifies non-base levels that may be grouped

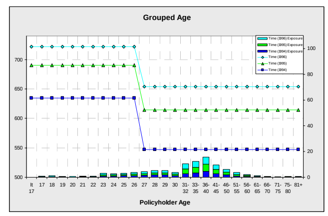
	Lt 17	17	18	19	20	21	22
Lt 17							
17	90.4						
18	85.6	308.9					
19	107.2	132.7	91.2				
20	92.7	995.9	255.1	161.6			
21	97.8	236.1	127.0	254.7	332.7		
22	95.4	362.2	163.9	199.5	620.3	685.0	
23	102.6	124.2	76.9	618.2	158.1	273.1	193.0
24	103.1	122.4	76.6	719.3	154.6	259.0	186.9
25	104.2	112.5	71.7	1,182.8	140.8	217.5	165.4
26	98.4	176.5	96.1	258.8	246.0	1,250.8	399.8
27	140.4	42.3	32.4	80.8	48.0	45.9	45.2
28	129.6	48.8	36.4	106.9	56.1	55.3	53.7
29	124.6	53.7	39.5	130.3	62.0	62.9	60.3
30	140.7	42.4	32.5	80.6	48.0	46.1	45.5
31-32	126.6	50.0	36.8	116.4	58.0	57.3	55.5
33-35	135.7	43.0	32.3	86.7	49.3	46.9	46.3
36-40	139.4	40.6	30.7	78.9	46.6	43.7	43.4

Build Models
Group Rating Levels

- Explore if “indicated” groupings are consistent over time or random parts of the data







- View of consistency without groupings
- View of consistency with groupings

CAS Predictive Modeling
Oct 2006

Build Models
Group Rating Levels

- Background
- Overall Strategy
- Modeling Steps
- 1. Get Data
- 2. Initial Sels
- 3. Test Error/Link
- 4. Preliminary Investigation
- 5. Build Models
- 6. Validate Models
- 7. Combine Models
- Summary

- Goodness of fit tests (e.g., Chi-Squared) can be used to determine the explanatory power of a variable
  - Null hypothesis is that the models with and without the grouping are the same

Chi-Squared

Model	With	Without
Deviance	8,906.4414	8,909.6226
Degrees of Freedom	18,469	18,475
Scale Parameter	0.4822	0.4823
Chi Square Test		78.6%

Score	H_0	Indicated Model
<5%	Reject	More Complex: Without Grouping
5%-30%	???	???
>30%	Accept	Simpler: With Grouping

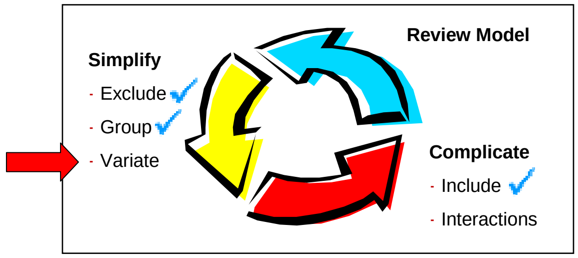
CAS Predictive Modeling

Oct 2006

• Background
• Overall Strategy
• Modeling Steps
1. Get Data
2. Initial Sels
3. Test Error/Link
4. Preliminary Investigation
5. Build Models
6. Validate Models
7. Combine Models
• Summary

Building the "Best" Model

- Modeling is an iterative process



- Variate: can the signal for a given variable be represented well by a curve?

EMB

CAS Predictive Modeling

Oct 2006

• Background
• Overall Strategy
• Modeling Steps
1. Get Data
2. Initial Sels
3. Test Error/Link
4. Preliminary Investigation
5. Build Models
6. Validate Models
7. Combine Models
• Summary

Build Models
Add Variates

- Curves can be applied to continuous variables, but not categorical variables
  - Continuous variables have a numerical relationship between the different levels

	Categorical	Continuous
Homeowners	Type of HO Alarm	Amount of Insurance
Auto	Vehicle Usage	Age of Driver
Commercial Lines	Occupation	Income
Retention	Gender	Premium change
Geography	Territory	Latitude/longitude

EMB

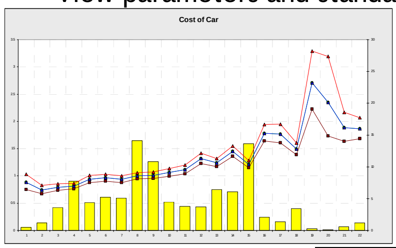
CAS Predictive Modeling

Oct 2006

Build Models

Add Variates

- View parameters and standard errors for sensibility of variate



- Variates can be very helpful at smoothing out non-sensible results
- Standard error of parameter differences can identify smooth progression of parameters

	Vehicle Group (1)	Vehicle Group (2)	Vehicle Group (3)	Vehicle Group (4)	Vehicle Group (5)	Vehicle Group (6)	Vehicle Group (7)	Vehicle Group (8)	Vehicle Group (9)	Vehicle Group (10)
Vehicle Group (1)										
Vehicle Group (2)	52.9									
Vehicle Group (3)	74.8	88.5								
Vehicle Group (4)	93.6	59.0	133.8							
Vehicle Group (5)	123.8	22.4	21.0	20.6						
Vehicle Group (6)	86.9	19.8	17.5	16.5	123.1					
Vehicle Group (7)	129.3	22.4	20.8	20.0	1,051.2	105.6				
Vehicle Group (8)	61.8	16.5	13.0	10.9	46.2	76.9	41.1			
Vehicle Group (9)	56.6	16.0	12.8	10.9	39.9	59.0	35.9	170.1		
Vehicle Group (10)	42.4	14.7	12.2	11.1	27.6	33.6	25.8	43.3	55.5	
Vehicle Group (11)	34.3	13.2	11.0	10.0	21.0	23.9	19.9	26.9	31.1	76.6
Vehicle Group (12)	20.1	9.4	7.5	6.7	10.7	11.2	10.2	10.8	11.6	16.7
Vehicle Group (13)	23.0	9.9	7.5	6.5	11.4	12.0	10.8	11.3	12.5	20.3
Vehicle Group (14)	15.9	7.7	5.7	4.8	7.5	7.5	7.0	6.7	7.2	10.2
Vehicle Group (15)	24.3	10.0	7.3	5.9	11.3	11.8	10.5	10.4	11.7	21.2

- Background
- Overall Strategy
- **Modeling Steps**
  1. Get Data
  2. Initial Sels
  3. Test Error/Link
  4. Preliminary Investigation
  5. **Build Models**
  6. Validate Models
  7. Combine Models
- Summary

EMB

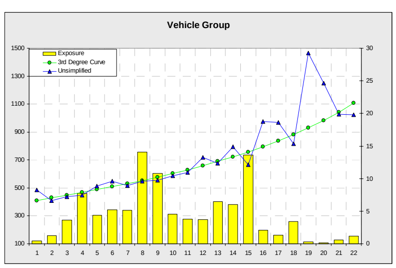
CAS Predictive Modeling

Oct 2006

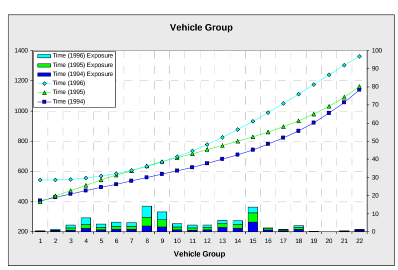
Build Models

Add Variates

- Check consistency of curve over time or random parts of the dataset



- After choosing the curve fits the data



- Check to see the consistency of that curve fit to different parts of the data

- Background
- Overall Strategy
- **Modeling Steps**
  1. Get Data
  2. Initial Sels
  3. Test Error/Link
  4. Preliminary Investigation
  5. **Build Models**
  6. Validate Models
  7. Combine Models
- Summary

EMB

CAS Predictive Modeling

Oct 2006

- Background
- Overall Strategy
- **Modeling Steps**
  1. Get Data
  2. Initial Sels
  3. Test Error/Link
  4. Preliminary Investigation
  5. **Build Models**
  6. Validate Models
  7. Combine Models
- Summary

EMB

Build Models

Add Variates

- Goodness of fit tests (e.g., Chi-Squared) can be used to determine the appropriateness of a variate
  - Null hypothesis is that the models with and without the variate are the same

Chi-Squared

Model	No Curve	Curve
Deviance	8,906.4460	9,020.2270
Degrees of Freedom	18,469	18,487
Scale Parameter	0.4822	0.4879
Chi Square Test		0.0%

Score	H_0	Indicated Model
<5%	Reject	More Complex: No Curve
5%-30%	???	???
>30%	Accept	Simpler: With Curve

CAS Predictive Modeling

Oct 2006

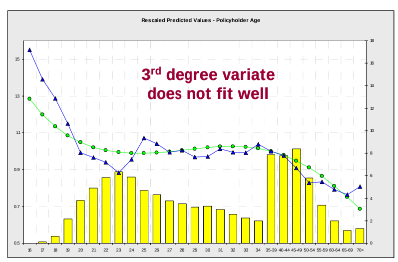
- Background
- Overall Strategy
- **Modeling Steps**
  1. Get Data
  2. Initial Sels
  3. Test Error/Link
  4. Preliminary Investigation
  5. **Build Models**
  6. Validate Models
  7. Combine Models
- Summary

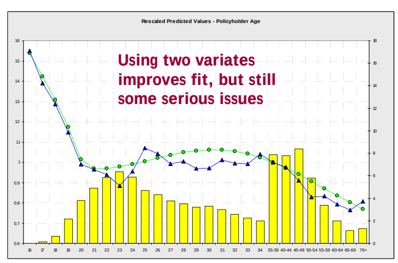
EMB

Build Models

Add Variates

- Variates tend not to perform as well with regards to Type III testing
- If variates are not fitting the data well, the modeler can increase the responsiveness
  - Increase the power of the variate
  - Create multiple variates
  - Use combination of groupings and variates
  - Fit splines





CAS Predictive Modeling
Oct 2006

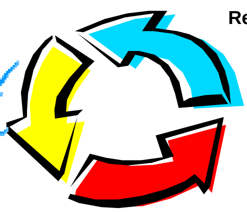
- Background
- Overall Strategy
- **Modeling Steps**
  1. Get Data
  2. Initial Sels
  3. Test Error/Link
  4. Preliminary Investigation
  5. **Build Models**
  6. Validate Models
  7. Combine Models
- Summary

EMB

Building the "Best" Model

- Modeling is an iterative process

Simplify
- Exclude ✓
- Group ✓
- Variate ✓



Review Model

Complicate
- Include ✓
- Interaction

- Interaction: does the effect of one variable vary by level of another variable?

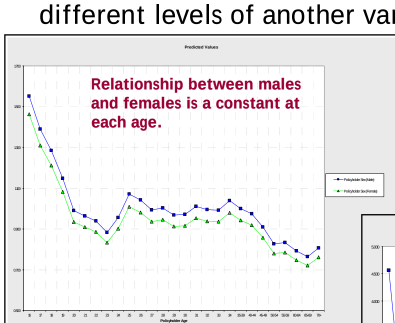
CAS Predictive Modeling
Oct 2006

- Background
- Overall Strategy
- **Modeling Steps**
  1. Get Data
  2. Initial Sels
  3. Test Error/Link
  4. Preliminary Investigation
  5. **Build Models**
  6. Validate Models
  7. Combine Models
- Summary

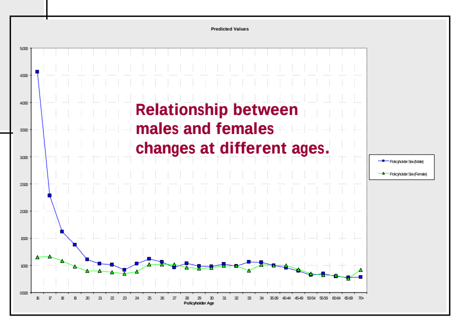
EMB

Build Models
Include Interactions

- Relationship of between levels of 1 variable may vary by different levels of another variable (e.g., response correlation)



Simple Model: Age + Gender



Full Interaction Model:

Age + Gender + Age.Gender

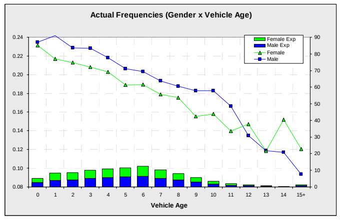
CAS Predictive Modeling
Oct 2006

Build Models
Identify Potential Interactions

- Background
- Overall Strategy
- Modeling Steps
  1. Get Data
  2. Initial Sels
  3. Test Error/Link
  4. Preliminary Investigation
  5. **Build Models**
  6. Validate Models
  7. Combine Models
- Summary

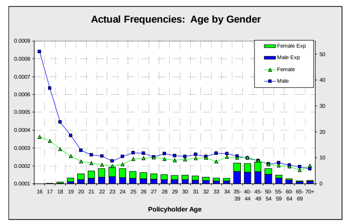
EMB

- Patterns of actual results will highlight potential interactions



- Actual frequencies support relationship between male and female is basically constant for each vehicle age

- Actual frequencies show relationship between male and female is very different for youthfuls and adults



CAS Predictive Modeling
Oct 2006

Build Models
Include Interactions

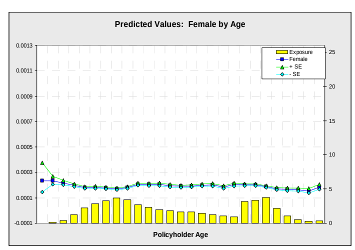
- Background
- Overall Strategy
- Modeling Steps
  1. Get Data
  2. Initial Sels
  3. Test Error/Link
  4. Preliminary Investigation
  5. **Build Models**
  6. Validate Models
  7. Combine Models
- Summary

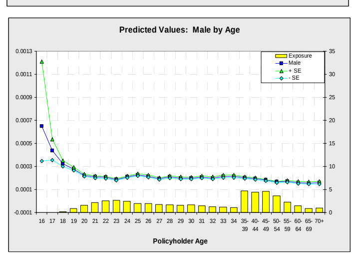
EMB

- View parameters and standard errors

Interaction Term	Value	Standard Error	Standard Error (%)	Weight
Female.16	-1.0235	0.78776	77.0	13,761
Female.17	-0.6174	0.24463	39.6	185,915
Female.18	-0.3981	0.11267	28.3	739,500
Female.19	-0.3382	0.07265	21.5	2,362,139
Female.20	-0.2112	0.06333	30.0	4,081,775
Female.21	-0.1384	0.05947	43.0	5,163,074
Female.22	-0.1467	0.05704	38.9	6,055,119
Female.23	-0.0782	0.05703	73.0	6,763,300
Female.24	-0.1536	0.05706	37.1	6,300,270
Female.25	-0.0972	0.05906	60.7	4,927,417
Female.26	-0.0431	0.06031	139.9	4,269,244
Female.27	0.0544	0.06364	116.9	3,672,472
Female.28	-0.0727	0.06477	89.1	3,438,810
Female.29	-0.0483	0.06761	140.0	2,970,306
Female.30	-0.0254	0.06693	263.3	3,027,278
Female.31	-0.0318	0.06849	215.1	2,724,535
Female.32	0.0033	0.07270	2,175.0	2,329,283
Female.33-35	-0.1597	0.07709	48.3	1,967,739
Female.36-39	-0.0376	0.07947	211.3	1,670,130
Female.40-44	0.0467	0.05185	111.1	6,166,191
Female.45-49	0.0297	0.05174	174.3	6,877,522
Female.50-54	0.0325	0.05973	183.8	3,957,251
Female.55-59	-0.0264	0.07412	281.0	1,998,839
Female.60-64	0.0228	0.09824	431.3	959,502
Female.65-69	-0.0168	0.13252	787.8	528,632
Female.70+	0.1593	0.12038	75.6	602,694

- In tabular format





- Graphically

CAS Predictive Modeling

Oct 2006

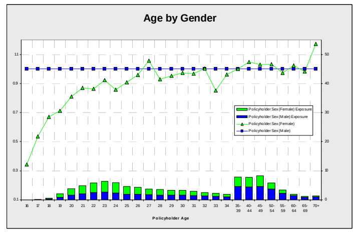
- Background
- Overall Strategy
- **Modeling Steps**
  1. Get Data
  2. Initial Sels
  3. Test Error/Link
  4. Preliminary Investigation
  5. **Build Models**
  6. Validate Models
  7. Combine Models
- Summary

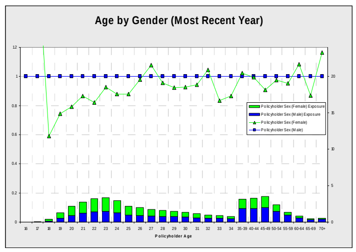
EMB

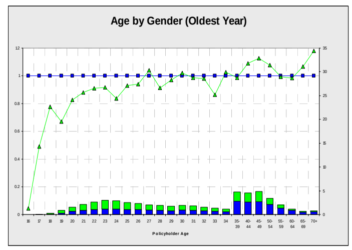
Build Models

Include Interactions

- Explore if “indicated” interaction is consistent over time or random parts of the data







CAS Predictive Modeling

Oct 2006

- Background
- Overall Strategy
- **Modeling Steps**
  1. Get Data
  2. Initial Sels
  3. Test Error/Link
  4. Preliminary Investigation
  5. **Build Models**
  6. Validate Models
  7. Combine Models
- Summary

EMB

Build Models

Include Interactions

- Goodness of fit tests (e.g., Chi-Squared) can be used to determine the explanatory power of an interaction
  - Null hypothesis is that the models with and without the interaction are the same

Chi-Squared

Model	Simple Model	W/ Interaction
Deviance	224,667.0000	224,771.0000
Degrees of Freedom	83	109
Scale Parameter	1.1615	1.1655
Chi Square Test		0.0%

Score	H_0	Indicated Model
<5%	Reject	More Complex: With Interaction
5%-30%	???	???
>30%	Accept	Simpler: Without Interaction

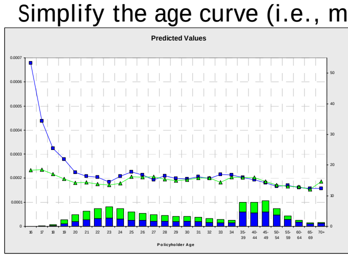
CAS Predictive Modeling
Oct 2006

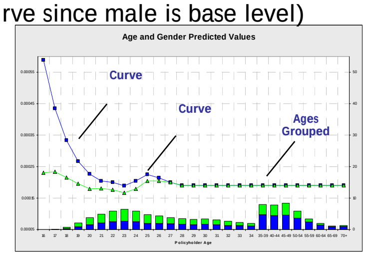
Build Models
Simplify Interactions

- Background
- Overall Strategy
- **Modeling Steps**
  1. Get Data
  2. Initial Sels
  3. Test Error/Link
  4. Preliminary Investigation
  5. **Build Models**
  6. Validate Models
  7. Combine Models
- Summary

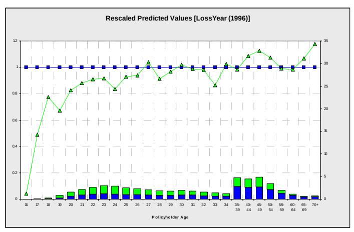
EMB

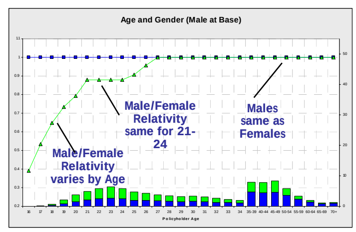
- Complex relationships can be simplified using curves, groups, etc.
  - Simplify the age curve (i.e., male curve since male is base level)





  - Simplify the relationship between males and females





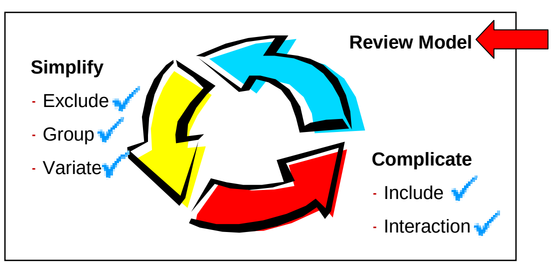
CAS Predictive Modeling
Oct 2006

Building the “Best” Model

- Background
- Overall Strategy
- **Modeling Steps**
  1. Get Data
  2. Initial Sels
  3. Test Error/Link
  4. Preliminary Investigation
  5. **Build Models**
  6. Validate Models
  7. Combine Models
- Summary

EMB

- Modeling is an iterative process



- Once models have been built, essential to validate the models

CAS Predictive Modeling

Oct 2006

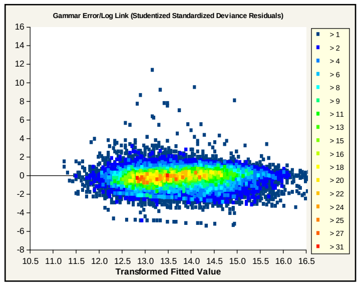
Validate Model

Residual Analysis

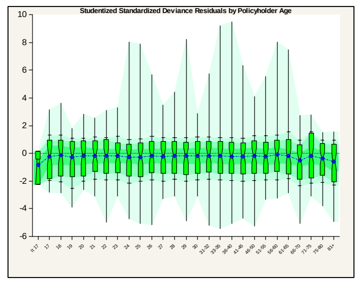
- Background
- Overall Strategy
- **Modeling Steps**
  1. Get Data
  2. Initial Sels
  3. Test Error/Link
  4. Preliminary Investigation
  5. Build Models
  6. **Validate Models**
  7. Combine Models
- Summary

EMB

- Re-check residuals to ensure appropriate shape



- Is the contour plot symmetric?
- Are fitted results reasonable?



- Does the Box-Whisker show symmetry across levels?

CAS Predictive Modeling

Oct 2006

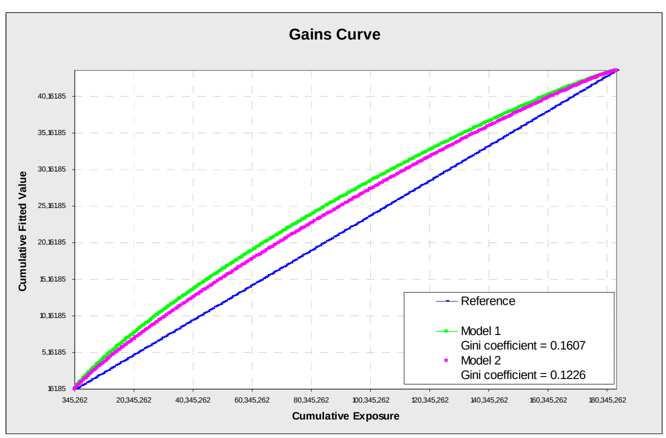
Validate Model

Gains Curves

- Background
- Overall Strategy
- **Modeling Steps**
  1. Get Data
  2. Initial Sels
  3. Test Error/Link
  4. Preliminary Investigation
  5. Build Models
  6. **Validate Models**
  7. Combine Models
- Summary

EMB

- Compare predictiveness of models



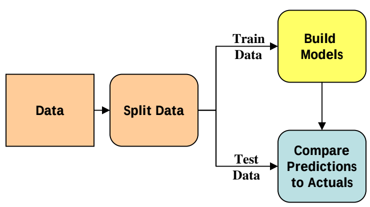
CAS Predictive Modeling
Oct 2006

Validate Model
Hold-out Samples

- Background
- Overall Strategy
- Modeling Steps
  1. Get Data
  2. Initial Sels
  3. Test Error/Link
  4. Preliminary Investigation
  5. Build Models
  6. Validate Models
  7. Combine Models
- Summary

- Hold-out samples are effective at validating model
  - Determine estimates based on part of dataset
  - Uses estimates to predict other part of dataset

Test/Training



- Larger companies may consider 3 splits
  1. Build models
  2. Fit parameters
  3. Validate models/parameters
- Smaller companies may consider a sampling approach

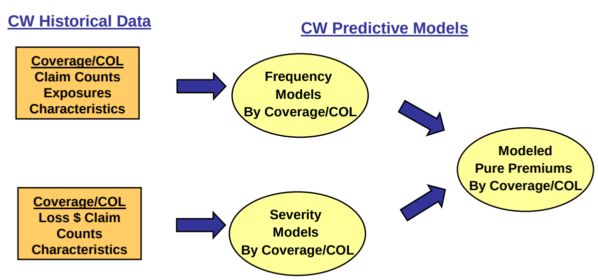
- Predictions should be close to actuals for populated cells

EMB

CAS Predictive Modeling
Oct 2006

Combine Predictive Models

- Background
- Overall Strategy
- Modeling Steps
  1. Get Data
  2. Initial Sels
  3. Test Error/Link
  4. Preliminary Investigation
  5. Build Models
  6. Validate Models
  7. Combine Models
- Summary



- Once signal determined, can implement business restrictions
  - Split variables into rating and underwriting
  - Incorporate parameter restrictions (e.g., cap relativities)
  - Incorporate structural restrictions (e.g., convert to mixed additive/multiplicative structure)

EMB

CAS Predictive Modeling
Oct 2006

- Background
- Overall Strategy
- Modeling Steps
  1. Get Data
  2. Initial Sels
  3. Test Error/Link
  4. Preliminary Investigation
  5. Build Models
  6. Validate Models
  7. Combine Models
- **Summary**

EMB

Summary

- GLMs can be a powerful tool modeling tool with significant advantages over traditional techniques
- Regardless of what is being modeled, the goal is to remove the "noise" and find the "signal" in the data
- When modeling risk, it is ideal to
  - Model frequency and severity separately
  - Model by coverage or cause of loss
  - Use all available data and worry about constraints later
- Modeling is a multi-step iterative process requiring the modeler to use statistical and practical tests and apply judgment

1 Get Clean Data → 2 Make Initial Selections → 3 Test Error Structure & Link Function → 4 Prelim Invest → 5 Build Model Structure → 6 Validate Models → 7 Combine Models

CAS Predictive Modeling
Oct 2006

EMB

Thanks for coming, if you would like a copy of these slides:

- Give me your name/email after the session
- Call me at 210.826.2878
- Email me at geoff.werner@embamerica.com

GLM III will cover:

- Testing the link function
- The Tweedie distribution
- Splines-theory and practice
- Reference models
- Aliasing/near-aliasing
- Combining models across claim types
- Restricted models
- Model validation