



EMB

# GLM II: Basic Modeling Strategy

CAS Predictive Modeling Seminar

Claudine Modlin, FCAS, MAAA

Senior Consultant

EMB America

October 6-7, 2008



# Basic Modeling Session

---

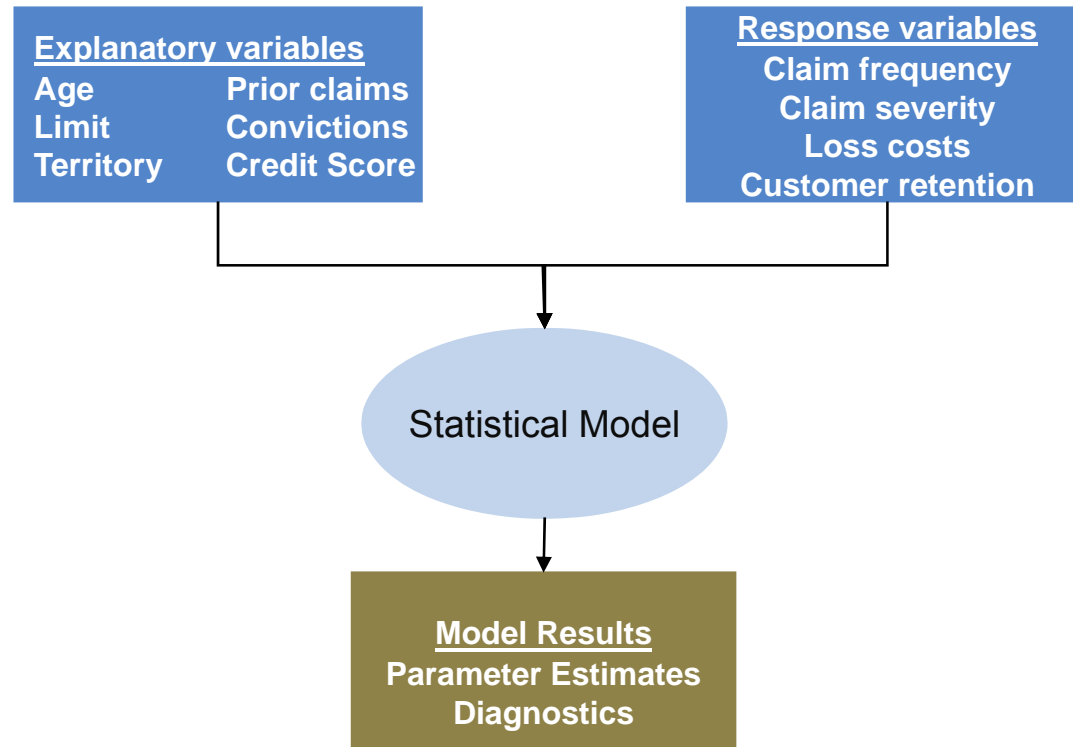
**PURPOSE:** To discuss basic strategies and techniques for building appropriate GLMs

## **OUTLINE**

- Background of GLMs
- Overall Modeling Strategy
- Basic Modeling Steps
  1. Get clean data
  2. Select an initial error structure, link function, and model structure
  3. Test error structure
  4. Review preliminary model effects
  5. Iterate models
  6. Validate final predictive models
  7. Combine models, if modeling frequency and severity
- Summary

# Purpose of Predictive Modeling

- To statistically measure the effect a series of explanatory variables has on an observed item, or response variable



*Same techniques apply regardless of what response variable is being modeled. This session will focus on claims modeling as it is the most common application of GLMs.*


# Background of Generalized Linear Models (GLMs)

---

Link function  
( $g=h^{-1}$ )

Model  
Structure

Error  
Structure

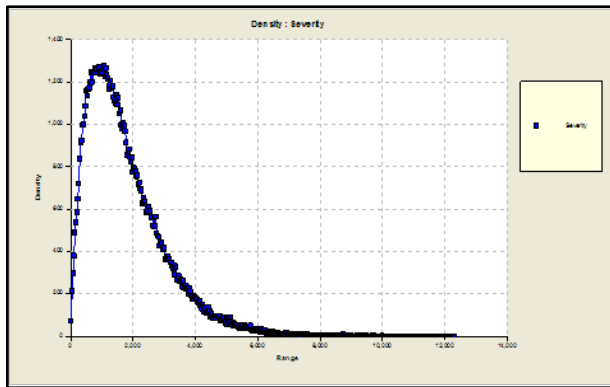

$$Y = h(X\beta + \xi) + \varepsilon$$

$Y = h(\text{Linear Combination of Factors}) + \text{Error}$

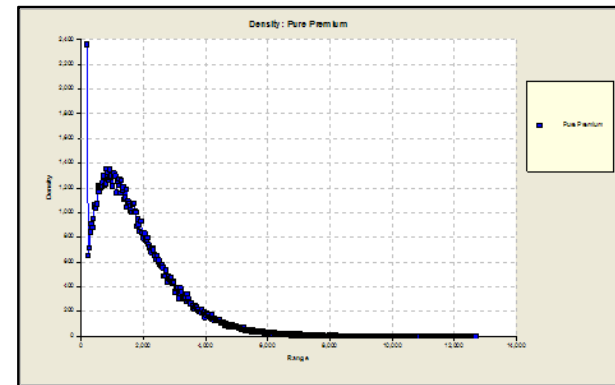
# GLM Building Blocks

$$y = h(\text{Linear Combination of Rating Factors}) + \text{Error}$$

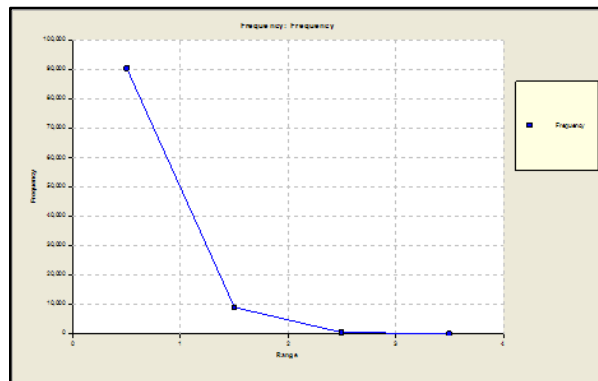
- Reflects the variability of the underlying process and can be any distribution within the exponential family, for example:



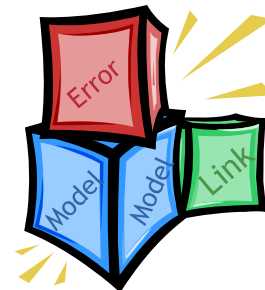
- Gamma consistent with severity modeling, may want to try Inverse Gaussian



- Tweedie consistent with pure premium modeling



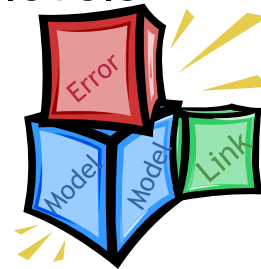
- Poisson consistent with frequency modeling



# GLM Building Blocks

$$y = h(\text{Linear Combination of Rating Factors}) + \text{Error}$$

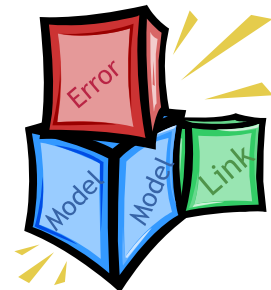
- Include variables that are predictive, exclude those that are not
  - e.g., gender may not have an effect on wind severity
- Simplify some factors
  - Some levels within a factor may be grouped together (e.g., 50-54 year olds)
  - A curve may replicate the factor effect (e.g., amount of insurance)
- Complicate model if the relationship between levels of one variable depends on another characteristic
  - e.g., the difference between males and females varies across levels of age



# GLM Building Blocks

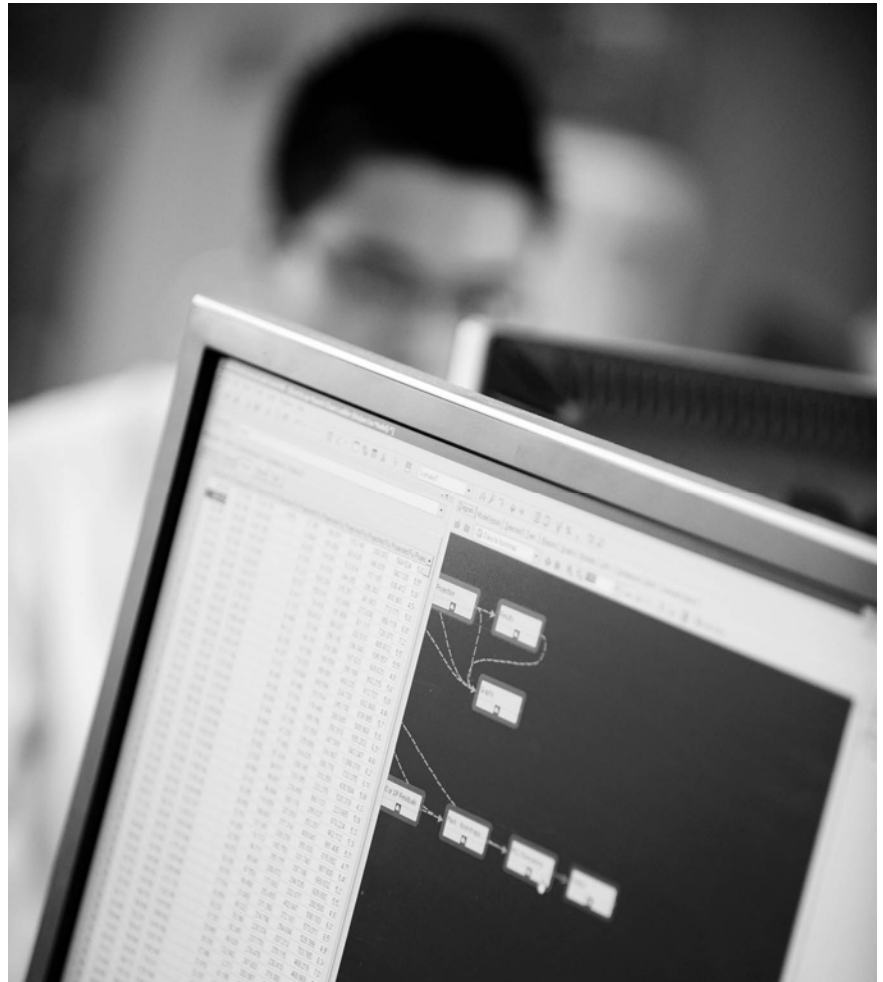
$$y = h(\text{Linear Combination of Rating Factors}) + \text{Error}$$

- Link function ( $g=h^{-1}$ ) chosen based on how the variables relate to one another to produce the best signal:
  - Log: variables relate multiplicatively (e.g., risk modeling)
  - Identity: variables relate additively (e.g., risk modeling)
  - Log it: retention or risk modelling



## Overall Modeling Strategy Questions

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



## Should You Model Loss Ratios?

---

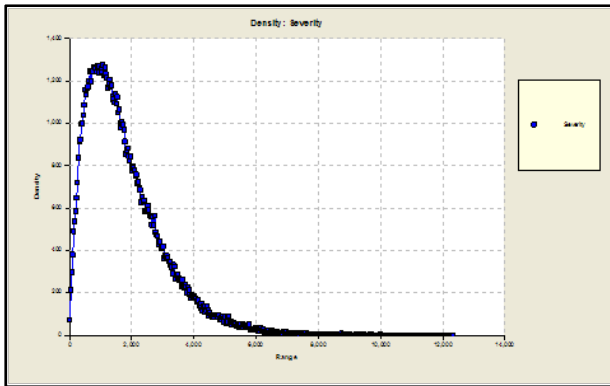
- Why some companies model loss ratios
  - May find it difficult to obtain exposures
  - Do not want to retrieve all rating variables, so assume using loss ratios will “adjust” for excluded variables
  - Habit formed when performing traditional analysis
- Theoretical and practical **disadvantages** to loss ratio modeling
  - On-level calculations
  - No defined error distribution
  - Difficult to distinguish noise from pattern
  - If changes made to premium, models cannot be re-used

- When modeling loss ratios, premiums should be put on-level to adjust for changes during or after the historical period
  - Rate change
  - 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 and data may not be available
- Depending on magnitude of the changes, not putting premiums on level can result in serious under- and over-predictions
- Pure premiums use exposures so this is a non-issue

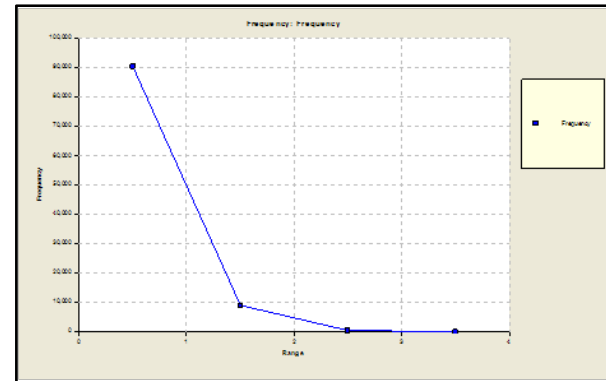
# Loss Ratio Modeling

## Defined Error Structure

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



**Gamma considered a standard for severity modeling**



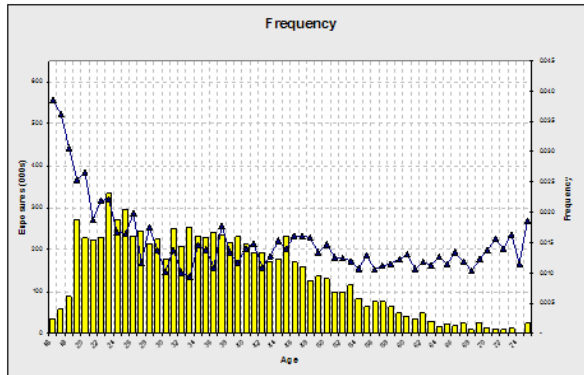
**Poisson considered a standard for frequency modeling**

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

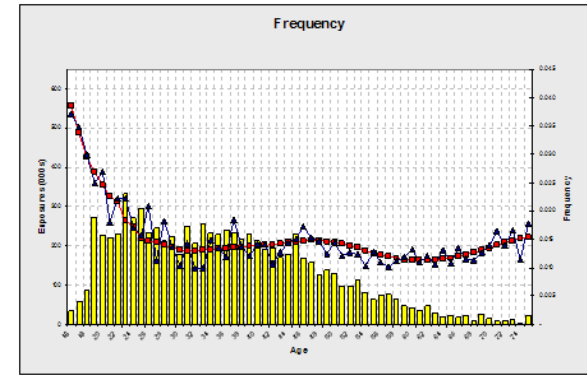
# Loss Ratio Modeling

## Distinguishing Patterns

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

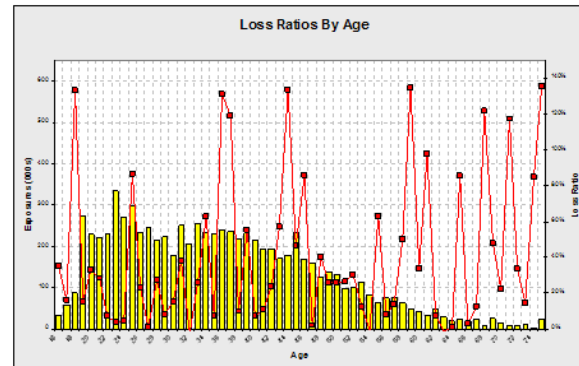


**Raw Frequency by Age of Driver**



**Smoothed Frequency by Age of Driver**

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



**Raw Loss Ratio by Age of Driver**

- Loss ratio modeling
  - Imperative that premiums be put on-level for each analysis
  - If a rate review results in changes
    - All of the loss ratios will change
    - The indicated loss ratio differentials may change as well
  - Models built in last review will be inappropriate
- Pure Premium modeling
  - Not necessary to put premiums on level
  - If a review results in changes
    - The frequencies, severities, pure premiums will not change
    - The indicated differentials will be unaffected
  - Models built in last review may still be appropriate
  - Can convert pure premium model to expected loss ratio model by offsetting by log of current annual premiums

# Granular or Combined Modeling?

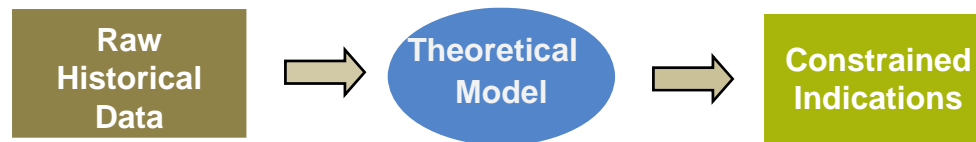
- Some tempted to model pure premiums or combined coverages/perils, presumably to save time
- As with traditional analysis (e.g., selecting loss trends), preferable to analyze at the granular level

Freq/Severity or Pure Premium	By-Peril or All Perils
Severity trends mask frequency signal	Highly 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

## 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



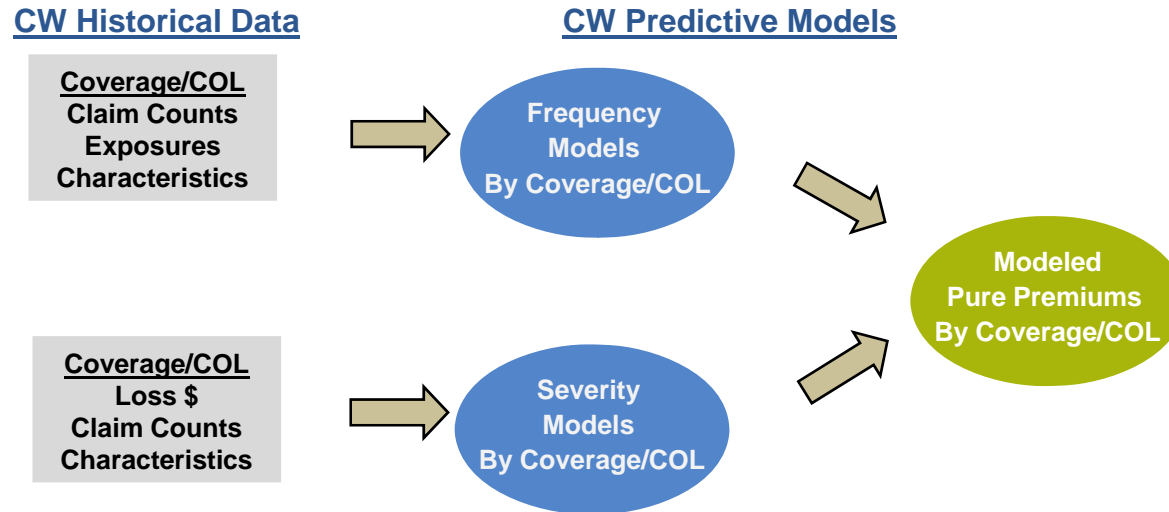
### Pure Modeling

- Use all data to remove “noise” and find signal
- Example, geodemographic data may be more predictive than current territories

### Constrained Modeling

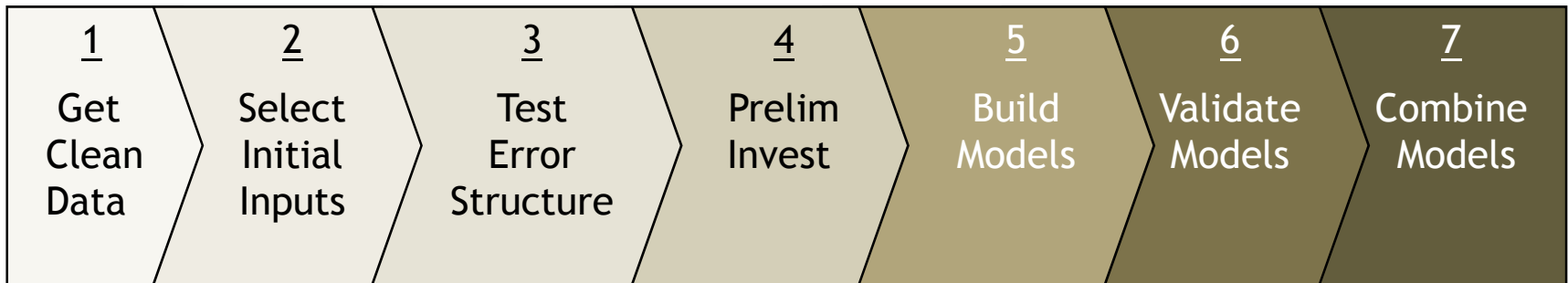
- Convert modeled results into usable indications
- Incorporate restrictions
  - IT
  - Regulatory
  - Competitive

# 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

# Basic Modeling Steps



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

# Get Clean Data

---

- Good project results start with good data (internal and external)
- Common data problems
  - Poor linkage between losses and policy characteristics
  - Data that does not reconcile to financial records
  - Negative or zero exposures; negative losses
  - Null records or bad data, especially for variables not used in rating
  - 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

# Initial Model 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	$\mu^0$
Claim Frequency	Log	Poisson	$\mu$
Claim Severity	Log	Gamma	$\mu^2$
Claim Severity	Log	Inverse Gaussian	$\mu^3$
Risk Premium	Log	Tweedie	$\mu^T$ where $1 < T < 2$
Retention Rate	Logit	Binomial	$\mu (1-\mu)$
Conversion Rate	Logit	Binomial	$\mu(1- \mu)$

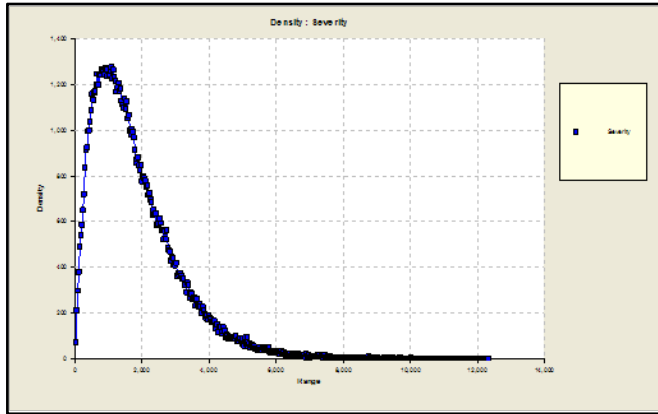
➤ Reasonable starting point for model structure

- All or all known important variables
- Prior model (last year or other related peril)
- Forward regression model

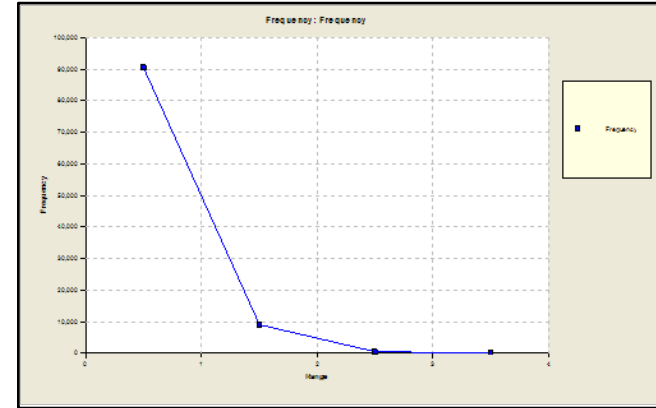
# Test Error Structure

## Distribution Analysis

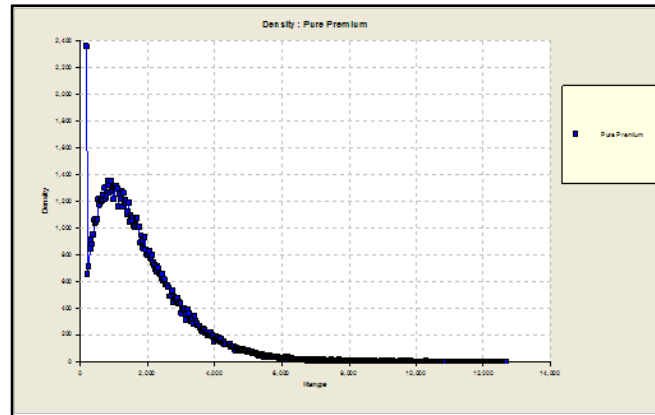
- Examine plots of the data (e.g., size of loss distribution)



- Consistent with gamma



- Consistent with Poisson

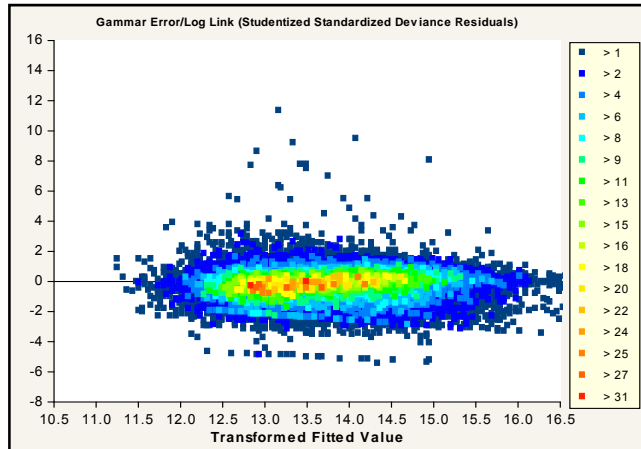


- Consistent with Tweedie

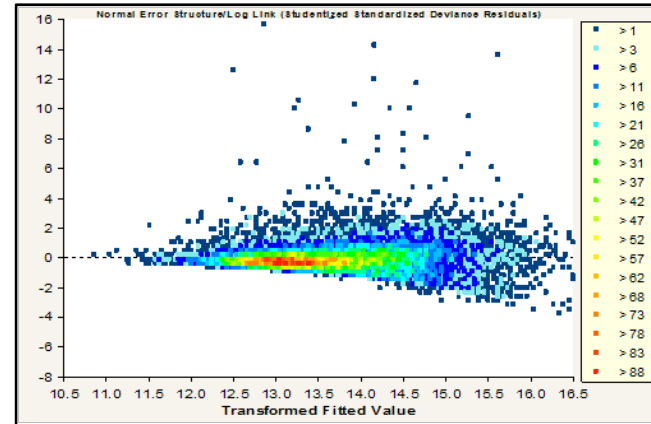
# Test Error Structure

## Residual Plots

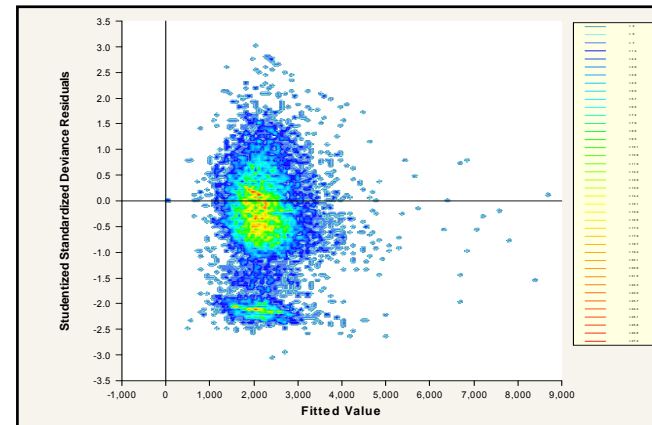
- Plot residuals to test selected error structure



- Elliptical pattern is ideal



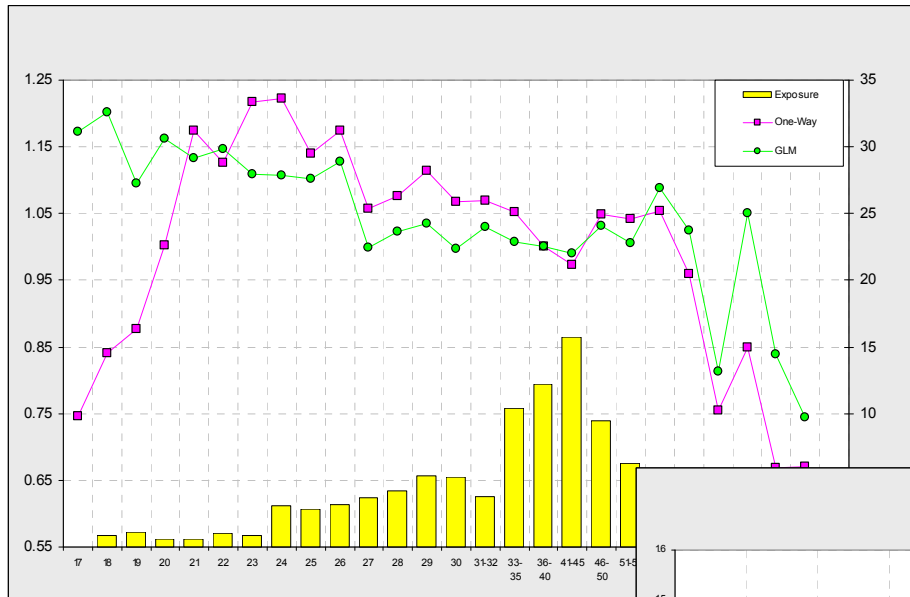
- Fanning out suggests power of variance function is too low



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

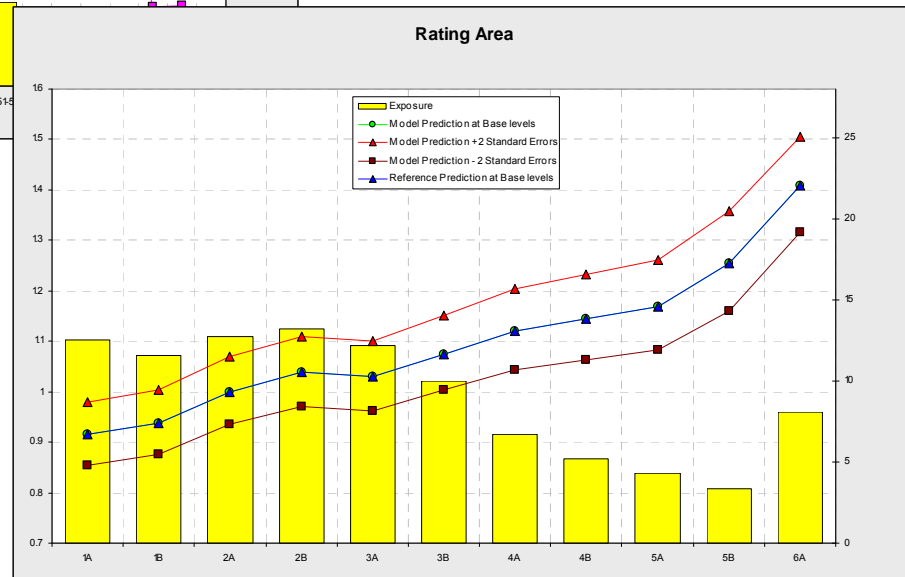
# Preliminary Investigation

➤ Simple graphs and traditional statistics provide “quick” feel



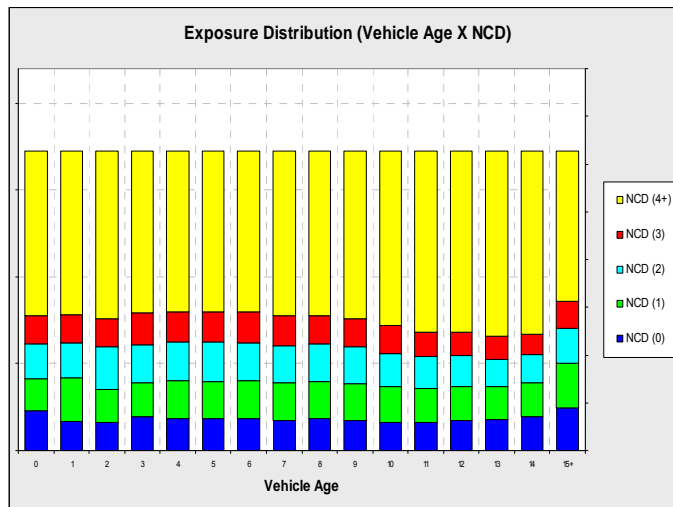
➤ Highlights what others within your company “know”

➤ Quickly highlights patterns in your data



# Preliminary Investigation

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

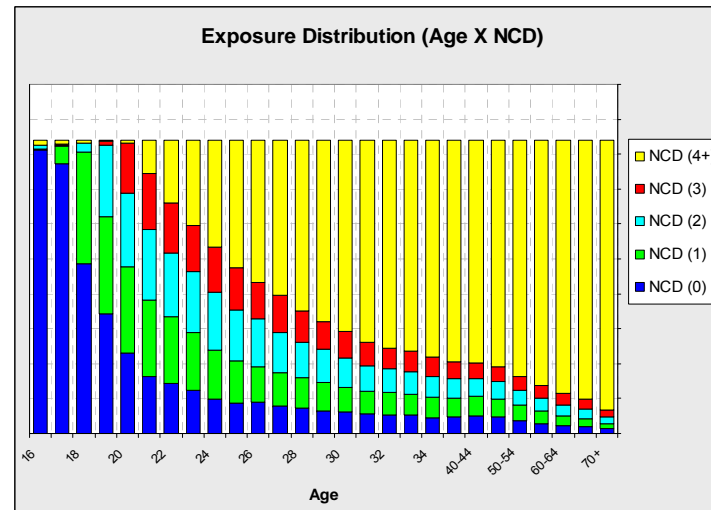


## Low Correlation (.025)

- Distribution of number of years claim-free is consistent across each vehicle age

## High Correlation (0.253)

- Older drivers are more likely to be claim-free



- Identifies independent variables that share predictive power

# Simple Model Parameter Notation

- Example: 2 rating variables (Age and Gender)

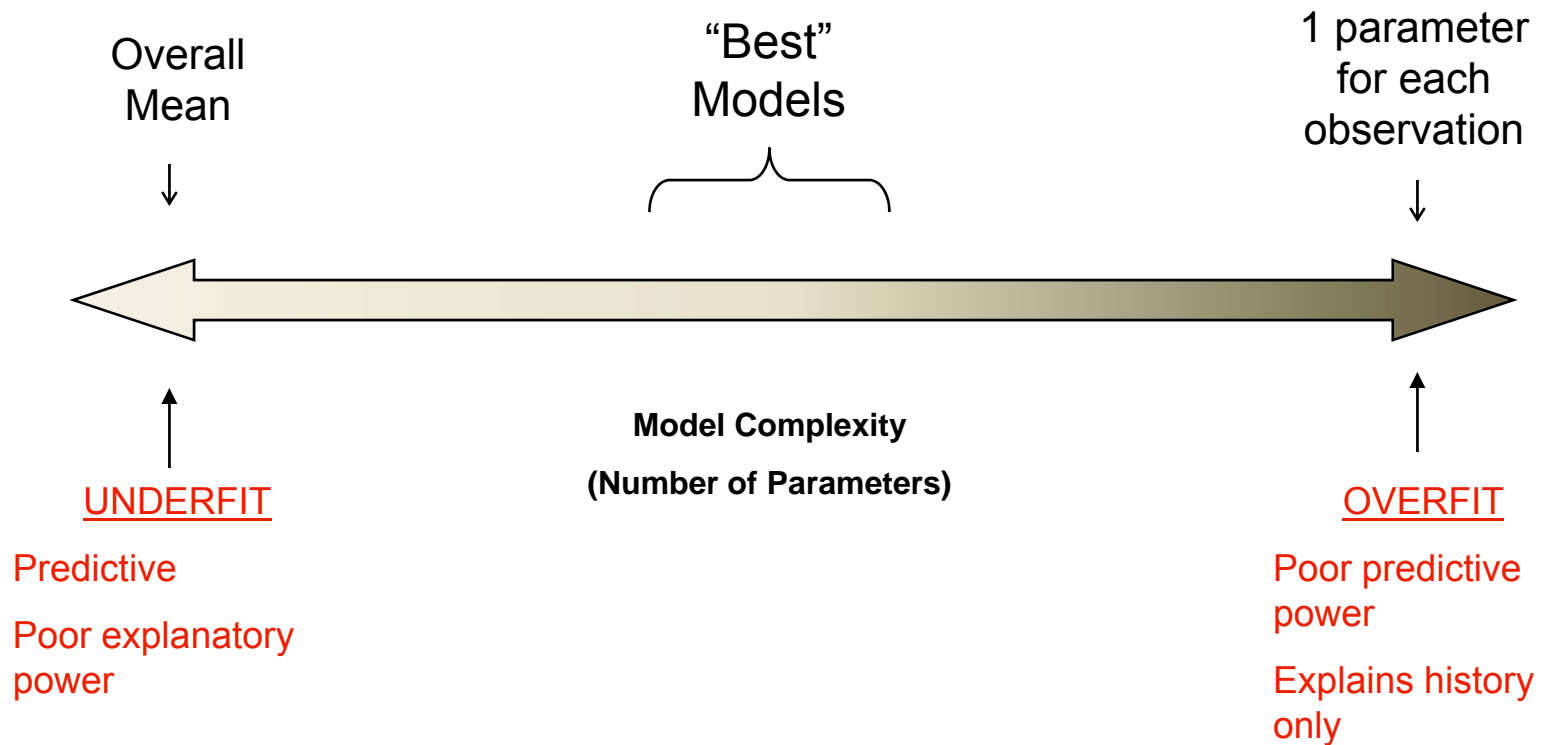
Simple Model: Age + Gender

	Male (Base)	Female
16	$\beta_0 + \beta_{16}$	$\beta_0 + \beta_{16} + \beta_F$
17	$\beta_0 + \beta_{17}$	$\beta_0 + \beta_{17} + \beta_F$
:	:	:
30 (Base)	$\beta_0$	$\beta_0 + \beta_F$
31	$\beta_0 + \beta_{31}$	$\beta_0 + \beta_{31} + \beta_F$
:	:	:
64	$\beta_0 + \beta_{64}$	$\beta_0 + \beta_{64} + \beta_F$
65+	$\beta_0 + \beta_{65+}$	$\beta_0 + \beta_{65+} + \beta_F$

- Log link:  $\text{Relativity}_{16,F} = \exp(\beta_0 + \beta_{16} + \beta_F) / \exp(\beta_0)$
- Identity link:  $\text{Additive}_{16,F} = (\beta_0 + \beta_{16} + \beta_F) - (\beta_0)$

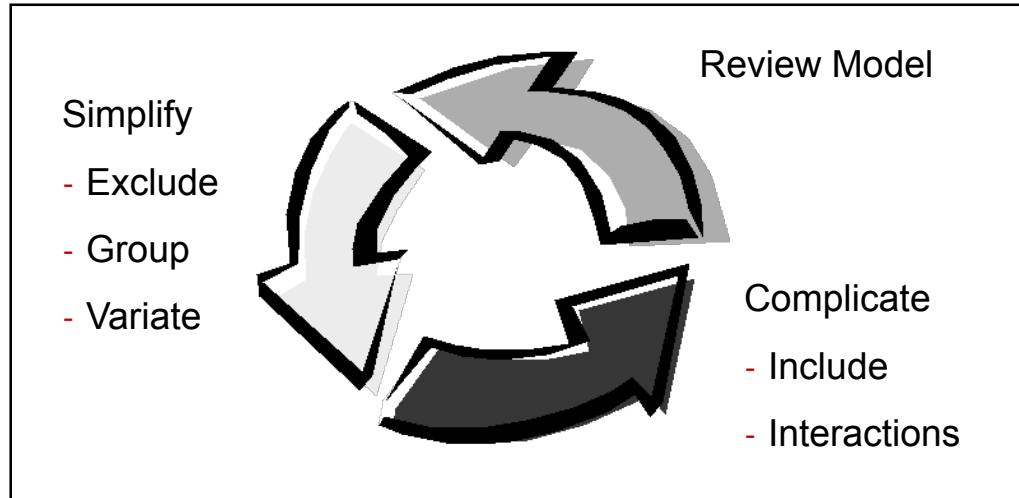
# Building the “Best” Model

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



# Building the “Best” Model

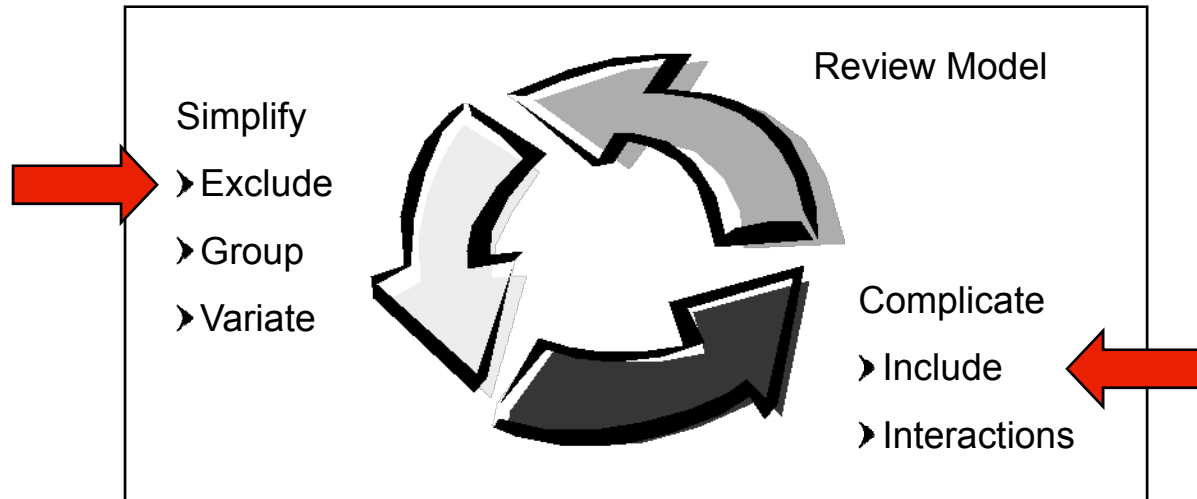
- Modeling is an iterative process



- 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 patterns make sense)

# Building the “Best” Model

- Modeling is an iterative process

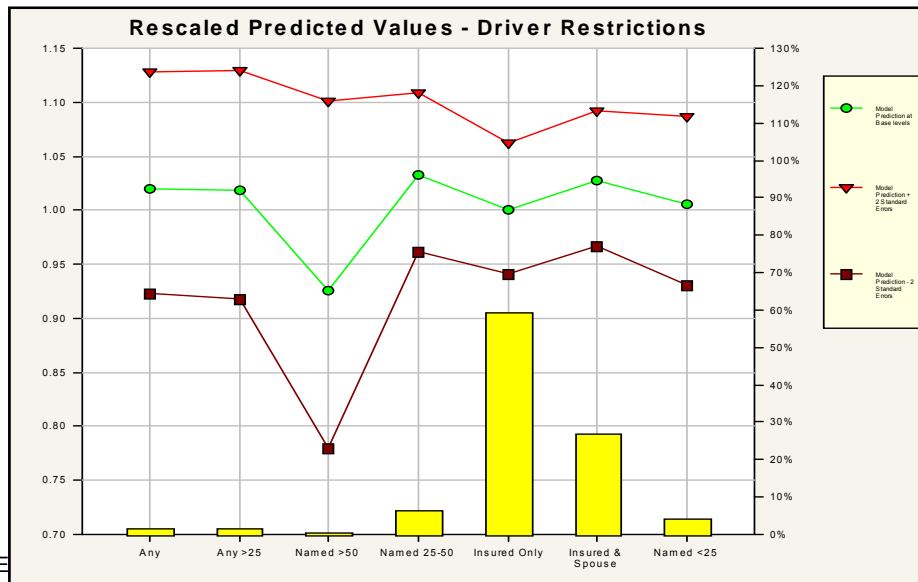


- Add/Exclude: does the independent variable have predictive power that warrants including it in the model?

- Parameter estimates (PEs) and standard errors (SEs) indicate strength and confidence in estimates

- If all PEs are roughly the same and/or have large SEs, the variable may not be predictive

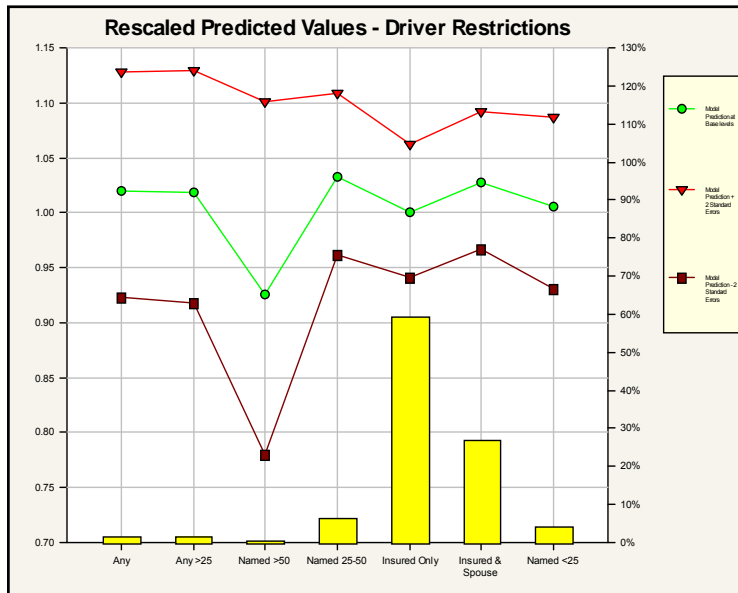
Name	Value	Standard Error	Standard Error (%)	Exp(Value)
Any	0.0174	0.04183	240.8	1.0175
Any>25	0.0212	0.04349	205.4	1.0214
Named >50	-0.0961	0.08120	84.5	0.9084
Named 25-50	0.0357	0.02194	61.4	1.0364
<b>Insured Only</b>				
Insured & Spouse	0.0255	0.01272	49.8	1.0259
Named <25	-0.0446	0.02663	59.7	0.9564



- Graph of PEs and SEs and “horizontal line test” identifies importance of a variable

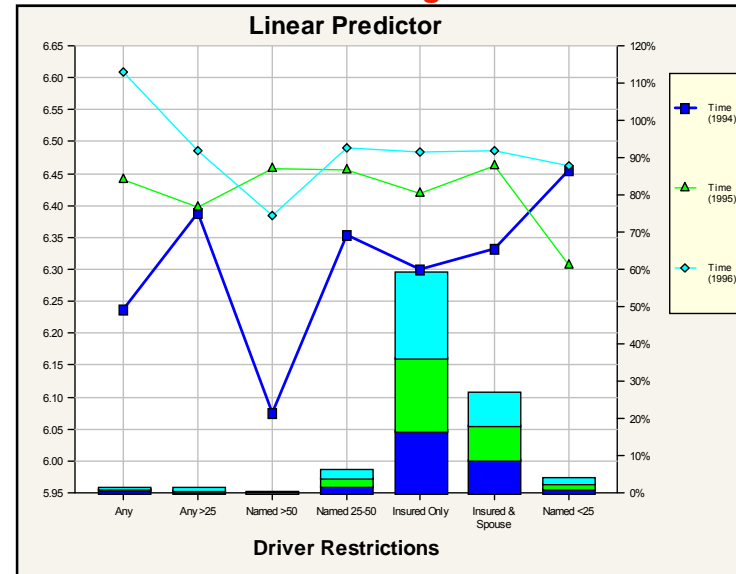
- Examine consistency over time or over random subsets

### Parameter/Standard Errors



- Main effects graph may show a questionable pattern

### Time Testing



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

# Build Models

## Include/Exclude Factors



- Statistical tests (e.g.,  $X^2$  or F-tests) can be used to determine the significance of a factor
  - Null hypothesis: models with and without a factor have the same statistical significance (alternative hypothesis suggests more complex model is better)

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%

Test result	$H_0$	Indicated Model
<5%	Reject	More Complex Model (i.e., include factor)
5%-30%	???	???
>30%	Accept	Simpler Model (i.e, exclude factor)

# Build Models

## Include/Exclude Factors



- Excluding a factor eliminates any variation due to that factor (e.g., remove gender)

Model: Age + Gender

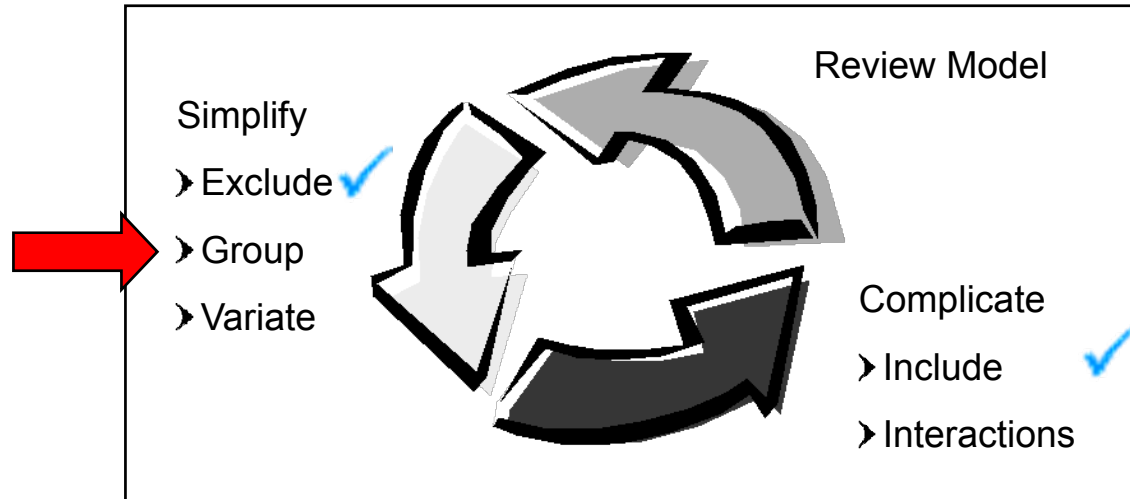
	Male (Base)	Female
16	$\beta_0 + \beta_{16}$	$\beta_0 + \beta_{16} + \beta_F$
17	$\beta_0 + \beta_{17}$	$\beta_0 + \beta_{17} + \beta_F$
:	:	:
30 (Base)	$\beta_0$	$\beta_0 + \beta_F$
31	$\beta_0 + \beta_{31}$	$\beta_0 + \beta_{31} + \beta_F$
:	:	:
64	$\beta_0 + \beta_{64}$	$\beta_0 + \beta_{64} + \beta_F$
65+	$\beta_0 + \beta_{65+}$	$\beta_0 + \beta_{65+} + \beta_F$

Model: Age

	Male (Base)	Female
16	$\beta_0 + \beta_{16}$	$\beta_0 + \beta_{16}$
17	$\beta_0 + \beta_{17}$	$\beta_0 + \beta_{17}$
:	:	:
30 (Base)	$\beta_0$	$\beta_0$
31	$\beta_0 + \beta_{31}$	$\beta_0 + \beta_{31}$
:	:	:
64	$\beta_0 + \beta_{64}$	$\beta_0 + \beta_{64}$
65+	$\beta_0 + \beta_{65+}$	$\beta_0 + \beta_{65+}$

# Building the “Best” Model

- Modeling is an iterative process



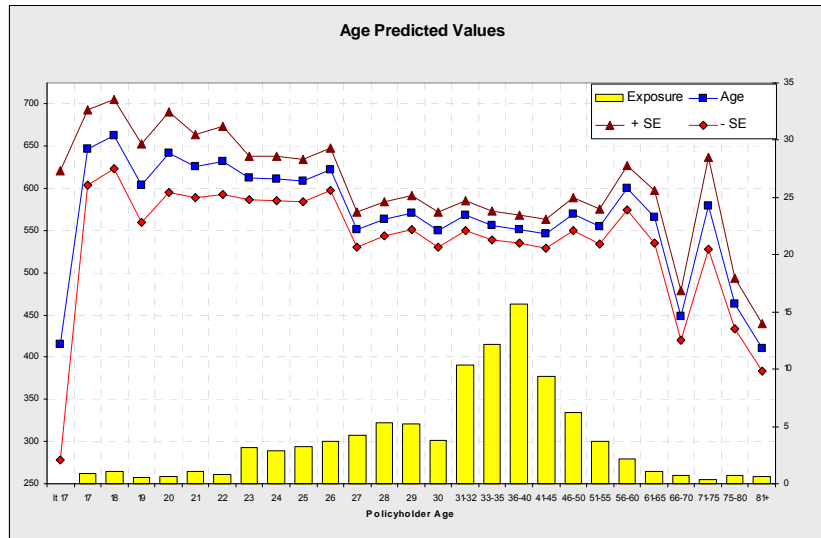
- Group: should some of the levels of a given variable be combined?

# Build Models

## Group Factor Levels



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



- Similar parameters or “plateaus” indicate potential groups
- Look for low volume

- Group levels with
  - Base level
  - Neighboring classes

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

# Build Models

## Group Factor 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

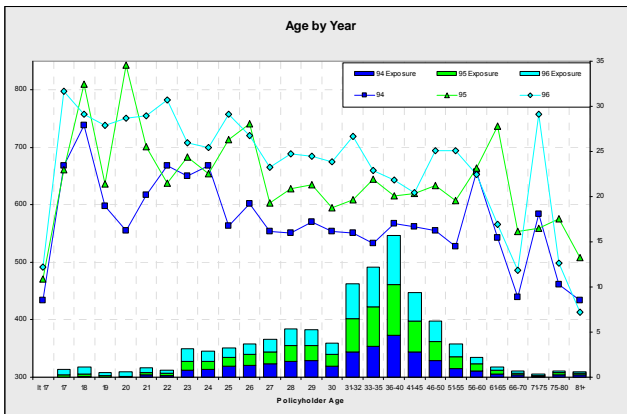
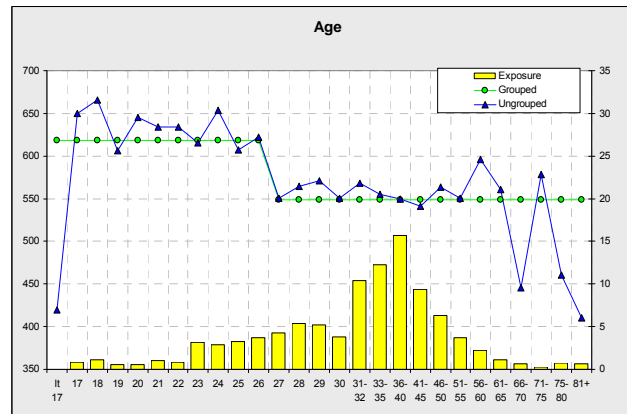
	Lt 17	17	18	19	20	21	22
Lt 17							
17	<b>90.4</b>						
18	<b>85.6</b>	<b>308.9</b>					
19	<b>107.2</b>	<b>132.7</b>	<b>91.2</b>				
20	<b>92.7</b>	<b>995.9</b>	<b>255.1</b>	<b>161.6</b>			
21	<b>97.8</b>	<b>236.1</b>	<b>127.0</b>	<b>254.7</b>	<b>332.7</b>		
22	<b>95.4</b>	<b>362.2</b>	<b>163.9</b>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

# Build Models

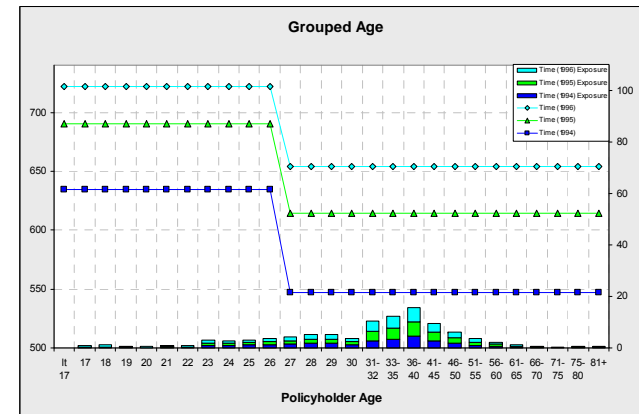
## Group Factor Levels



- Explore if proposed groupings are consistent over time or random subsets of the data



- Consistency without groupings



- Consistency with groupings

# Build Models

## Group Factor Levels



- Statistical tests (e.g.,  $\chi^2$  or F-tests) can be used to determine the statistical significance of a re-grouped variable
- Null hypothesis is that the original model and model with factor re-grouped have the same statistical significance

Score	H <sub>0</sub>	Indicated Model
<5%	Reject	More Complex: Without Grouping
5%-30%	???	???
>30%	Accept	Simpler: With Grouping

# Build Models

## Group Factor Levels



- Grouping forces multiple levels within a factor to have the same parameter estimates and standard errors

Model: Age + Gender

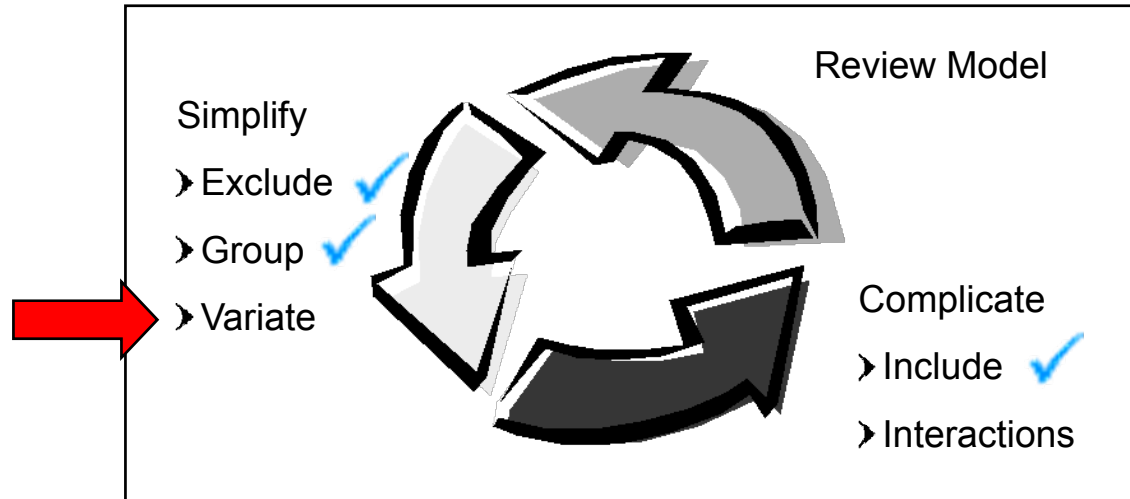
	Male (Base)	Female
16	$\beta_0 + \beta_{16}$	$\beta_0 + \beta_{16} + \beta_F$
17	$\beta_0 + \beta_{17}$	$\beta_0 + \beta_{17} + \beta_F$
:	:	:
30 (Base)	$\beta_0$	$\beta_0 + \beta_F$
31	$\beta_0 + \beta_{31}$	$\beta_0 + \beta_{31} + \beta_F$
:	:	:
64	$\beta_0 + \beta_{64}$	$\beta_0 + \beta_{64} + \beta_F$
65+	$\beta_0 + \beta_{65+}$	$\beta_0 + \beta_{65+} + \beta_F$

Model: **Grouped Age** + Gender

	Male (Base)	Female
16	$\beta_0 + \beta_{16-24}$	$\beta_0 + \beta_{16-24} + \beta_F$
17	$\beta_0 + \beta_{16-24}$	$\beta_0 + \beta_{16-24} + \beta_F$
:	:	:
30 (Base)	$\beta_0$	$\beta_0 + \beta_F$
31	$\beta_0 + \beta_{31+}$	$\beta_0 + \beta_{31+} + \beta_F$
:	:	:
64	$\beta_0 + \beta_{31+}$	$\beta_0 + \beta_{31+} + \beta_F$
65+	$\beta_0 + \beta_{31+}$	$\beta_0 + \beta_{31+} + \beta_F$

# Building the “Best” Model

- Modeling is an iterative process



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

# Build Models

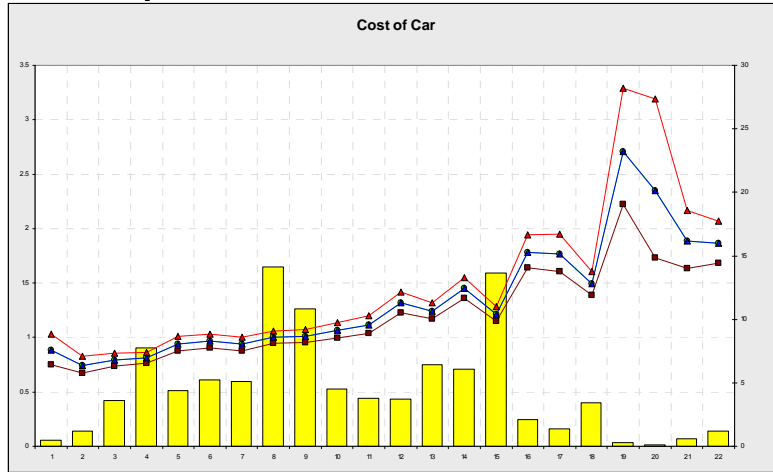
## Incorporate Variates



- Curves can be fit to continuous variables, but not discrete (a.k.a. categorical) variables
  - Levels of a continuous variable have a natural, numerical relationship

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

- View parameters and standard errors for sensibility of variate

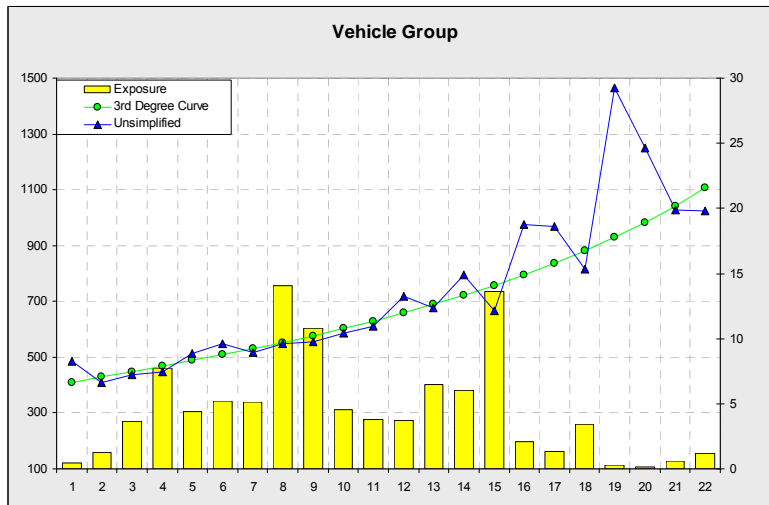


- Variates can be very helpful at smoothing out non-sensible results

- Standard errors of parameter differences can identify smooth progression of parameters

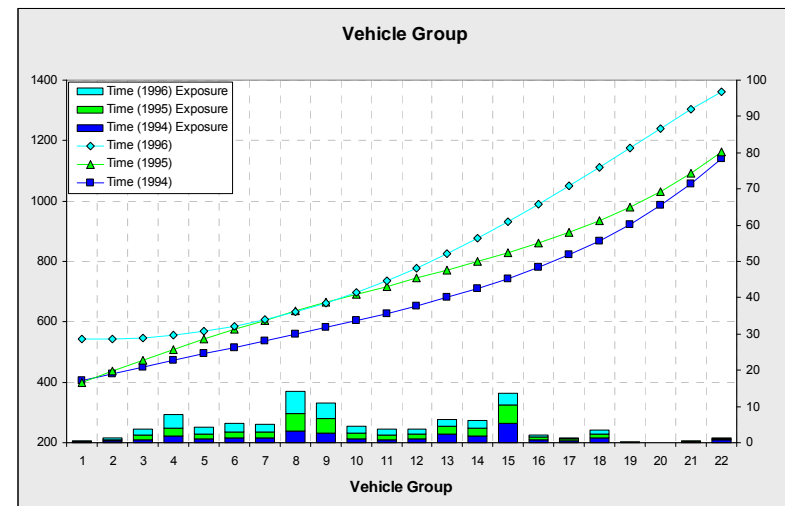
	Vehicle Group (1)	Vehicle Group (2)	Vehicle Group (3)	Vehicle Group (4)	Vehicle Group (5)	Vehicle Group (6)	Vehicle Group (7)	Vehicle Group (8)	Vehicle Group (9)	Vehicle Group (10)
Vehicle Group (1)										
Vehicle Group (2)	52.9									
Vehicle Group (3)	74.8	88.5								
Vehicle Group (4)	93.6	59.0	133.8							
Vehicle Group (5)	123.8	224	210	206						
Vehicle Group (6)	86.9	98	17.5	6.5	123.1					
Vehicle Group (7)	129.3	224	20.8	20.0	1,051.2	105.6				
Vehicle Group (8)	61.8	6.5	13.0	0.9	46.2	76.9	411			
Vehicle Group (9)	56.6	6.0	12.8	0.9	39.9	59.0	35.9	170.1		
Vehicle Group (10)	424	14.7	12.2	11.1	27.6	336	25.8	43.3	55.5	
Vehicle Group (11)	34.3	13.2	11.0	10.0	21.0	239	19.9	26.9	31.1	76.6
Vehicle Group (12)	20.1	94	7.5	6.7	10.7	112	10.2	10.8	116	6.7
Vehicle Group (13)	23.0	9.9	7.5	6.5	114	120	10.8	11.3	12.5	20.3
Vehicle Group (14)	6.9	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	118	10.5	10.4	11.7	21.2

- Check consistency of curve over time or random subsets of the data



- After choosing the curve

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



# Build Models

## Incorporate Variates



- Statistical tests (e.g.,  $\chi^2$  or F-tests) 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

# Build Models

## Incorporate Variates



- Levels of a continuous variable can be replaced with a curve (e.g., use a second degree polynomial)

Model: Age + Gender

	Male (Base)	Female
16	$\beta_0 + \beta_{16}$	$\beta_0 + \beta_{16} + \beta_F$
17	$\beta_0 + \beta_{17}$	$\beta_0 + \beta_{17} + \beta_F$
:	:	:
30 (Base)	$\beta_0$	$\beta_0 + \beta_F$
31	$\beta_0 + \beta_{31}$	$\beta_0 + \beta_{31} + \beta_F$
:	:	:
64	$\beta_0 + \beta_{64}$	$\beta_0 + \beta_{64} + \beta_F$
65+	$\beta_0 + \beta_{65+}$	$\beta_0 + \beta_{65+} + \beta_F$

Model: **Age Curve** + Gender

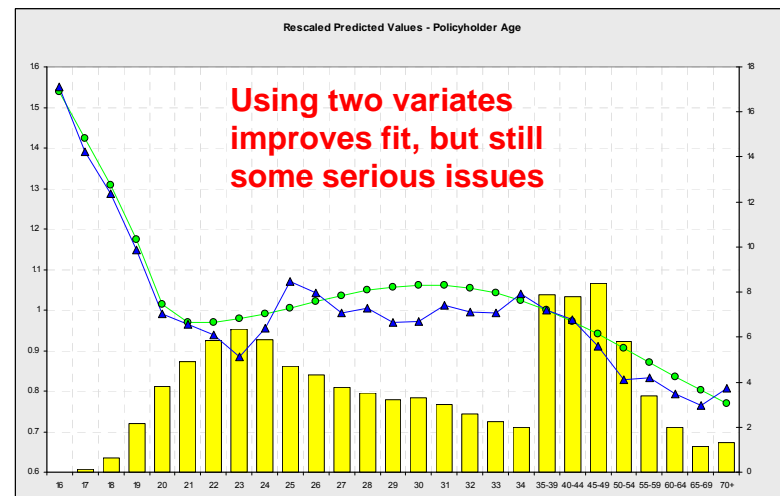
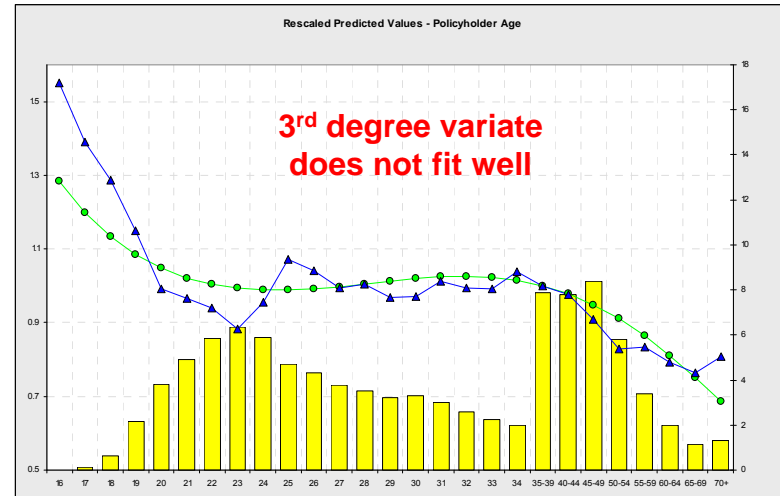
	Male (Base)	Female
16	$\beta_0 + \alpha_1 16 + \alpha_2 16^2$	$\beta_0 + \alpha_1 16 + \alpha_2 16^2 + \beta_F$
17	$\beta_0 + \alpha_1 17 + \alpha_2 17^2$	$\beta_0 + \alpha_1 17 + \alpha_2 17^2 + \beta_F$
:	:	:
30 (Base)	$\beta_0$	$\beta_0 + \beta_F$
31	$\beta_0 + \alpha_1 31 + \alpha_2 31^2$	$\beta_0 + \alpha_1 31 + \alpha_2 31^2 + \beta_F$
:	:	:
64	$\beta_0 + \alpha_1 64 + \alpha_2 64^2$	$\beta_0 + \alpha_1 64 + \alpha_2 64^2 + \beta_F$
65+	$\beta_0 + \alpha_1 70 + \alpha_2 70^2$	$\beta_0 + \alpha_1 70 + \alpha_2 70^2 + \beta_F$

# Build Models

## Incorporate Variates

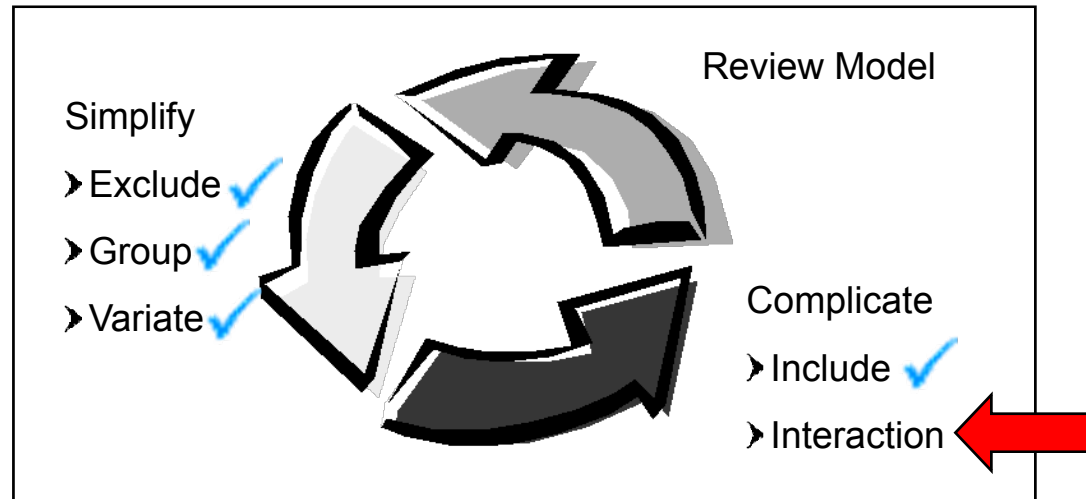


- Variates tend not to perform as well with regards to Type III testing
- If variates are not fitting the data well, the modeler can increase the responsiveness
  - Increase the power of the polynomial
  - Create multiple variates
  - Use combination of groupings and variates
  - Fit splines



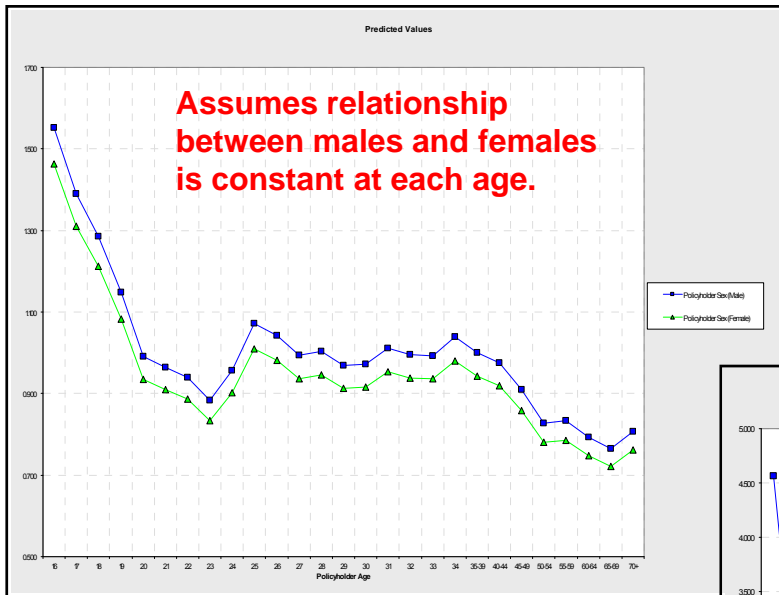
# Building the “Best” Model

- Modeling is an iterative process



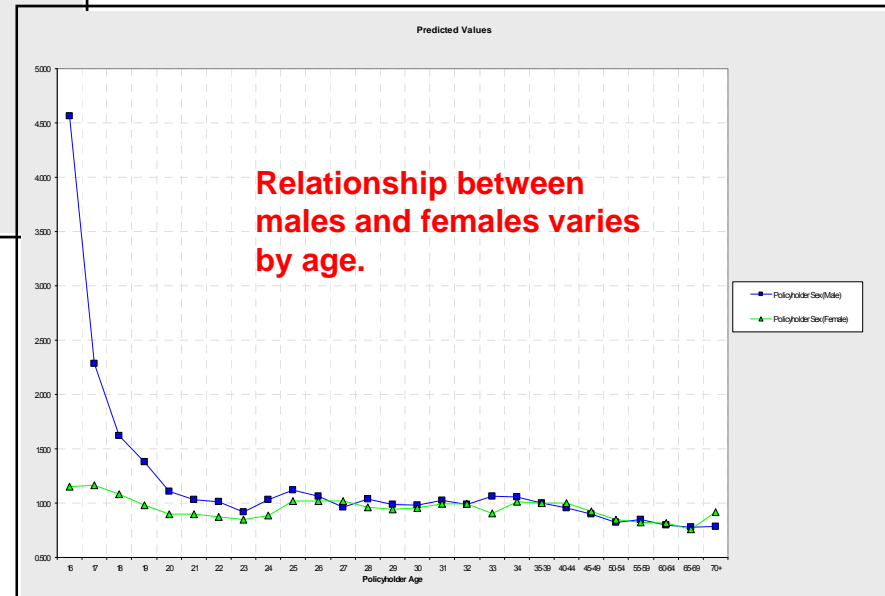
- Interaction: does the effect of one variable vary by level of another variable?

- Relationship between levels of one variable may vary by levels of another variable (i.e., response correlation)



Simple Model: Age + Gender

Full Interaction Model:  
Age + Gender + Age.Gender

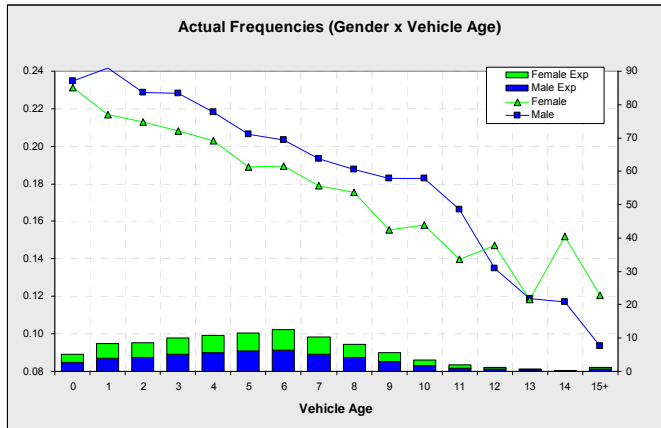


# Build Models

## Identify Potential Interactions

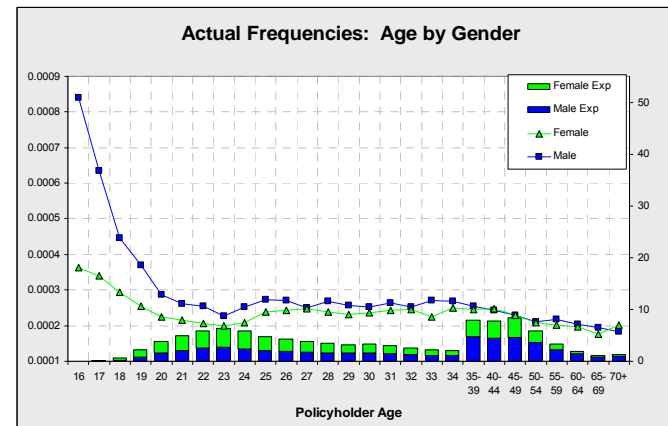


- ▶ Patterns of actual results highlight potential interactions



- ▶ 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 youth and adults



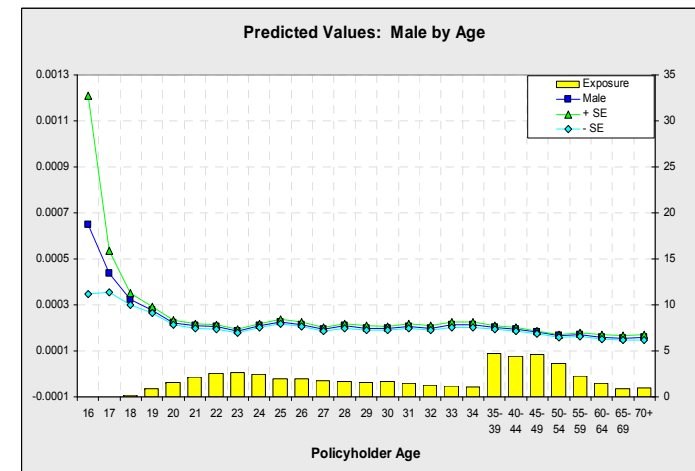
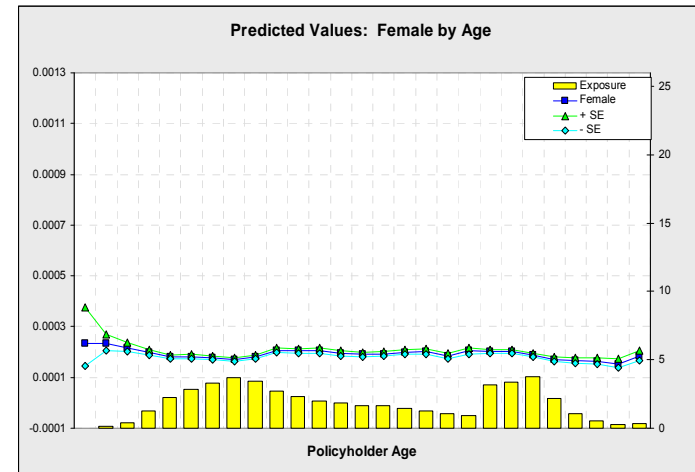
# Build Models

## Include Interactions



### ► 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

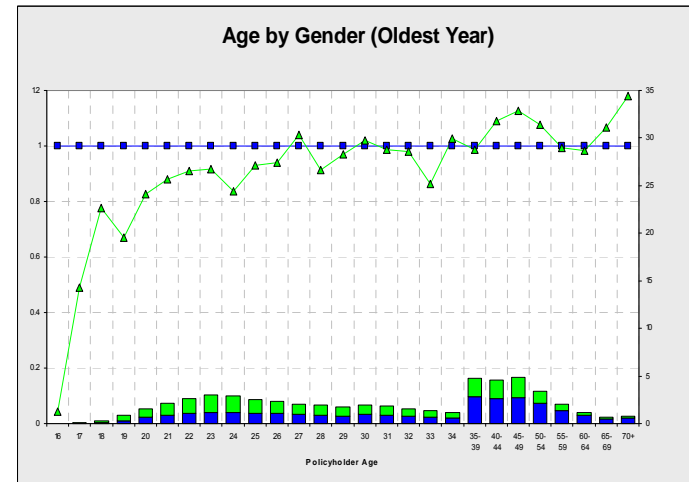
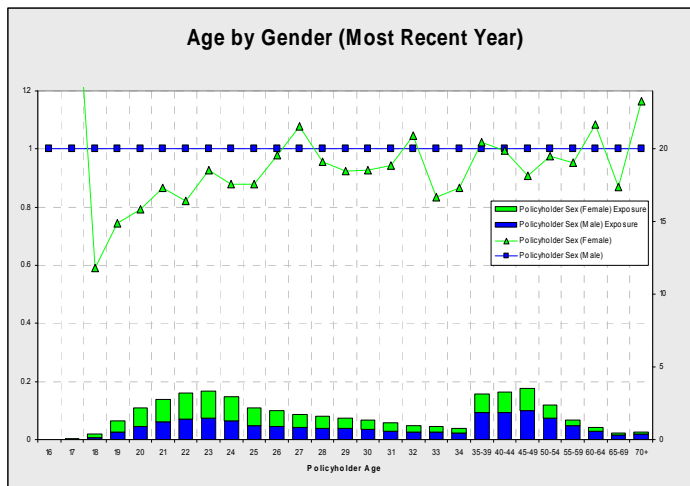
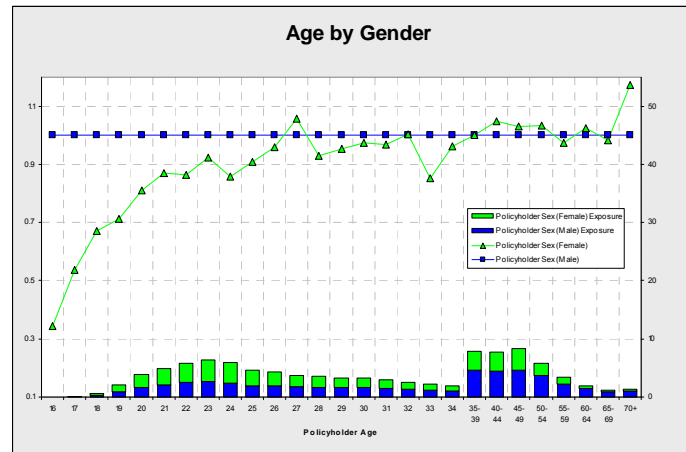
### ► Graphically

# Build Models

## Include Interactions



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



# Build Models

## Include Interactions



- Statistical tests (e.g.,  $\chi^2$  or F-tests) 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

# Build Models

## Include Interactions



- A full interaction allows the relationship between the levels of one variable to vary for each level of another variable

Model: Age + Gender

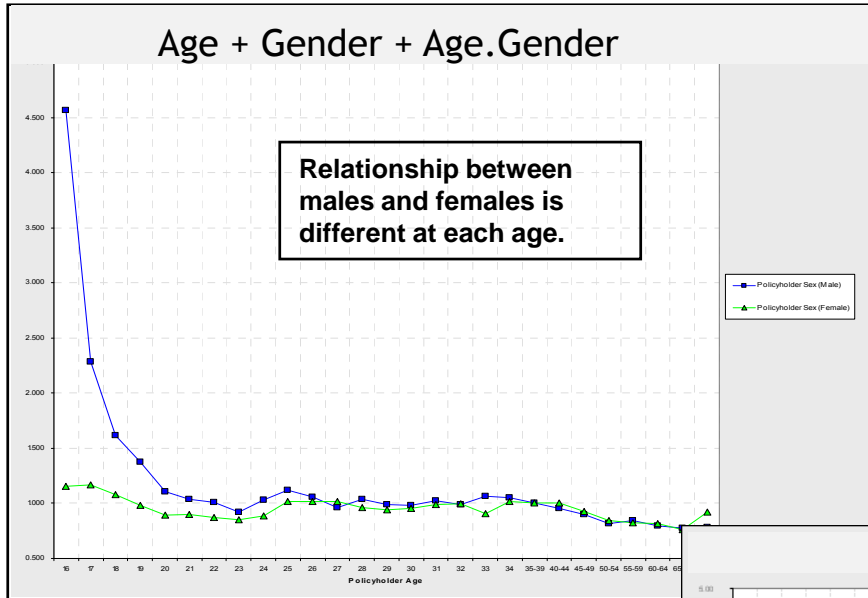
	Male (Base)	Female
16	$\beta_0 + \beta_{16}$	$\beta_0 + \beta_{16} + \beta_F$
17	$\beta_0 + \beta_{17}$	$\beta_0 + \beta_{17} + \beta_F$
:	:	:
30 (Base)	$\beta_0$	$\beta_0 + \beta_F$
31	$\beta_0 + \beta_{31}$	$\beta_0 + \beta_{31} + \beta_F$
:	:	:
64	$\beta_0 + \beta_{64}$	$\beta_0 + \beta_{64} + \beta_F$
65+	$\beta_0 + \beta_{65+}$	$\beta_0 + \beta_{65+} + \beta_F$

Model: Age + Gender + **Age.Gender**

	Male (Base)	Female
16	$\beta_0 + \beta_{16}$	$\beta_0 + \beta_{16} + \beta_F + \beta_{16,F}$
17	$\beta_0 + \beta_{17}$	$\beta_0 + \beta_{17} + \beta_F + \beta_{17,F}$
:	:	:
30 (Base)	$\beta_0$	$\beta_0 + \beta_F$
31	$\beta_0 + \beta_{31}$	$\beta_0 + \beta_{31} + \beta_F + \beta_{31,F}$
:	:	:
64	$\beta_0 + \beta_{64}$	$\beta_0 + \beta_{64} + \beta_F + \beta_{64,F}$
65+	$\beta_0 + \beta_{65+}$	$\beta_0 + \beta_{65+} + \beta_F + \beta_{65+,F}$

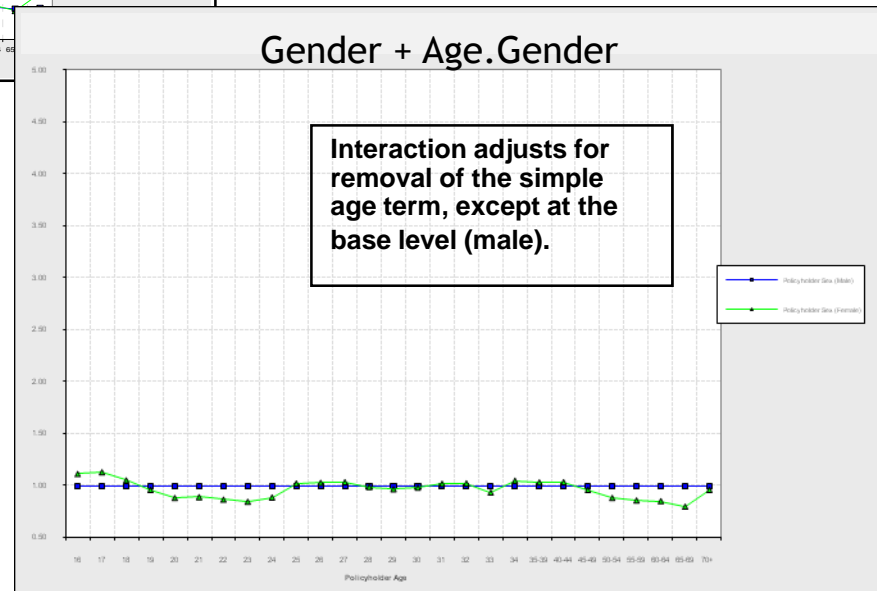
# Build Models

## Include Partial Interactions

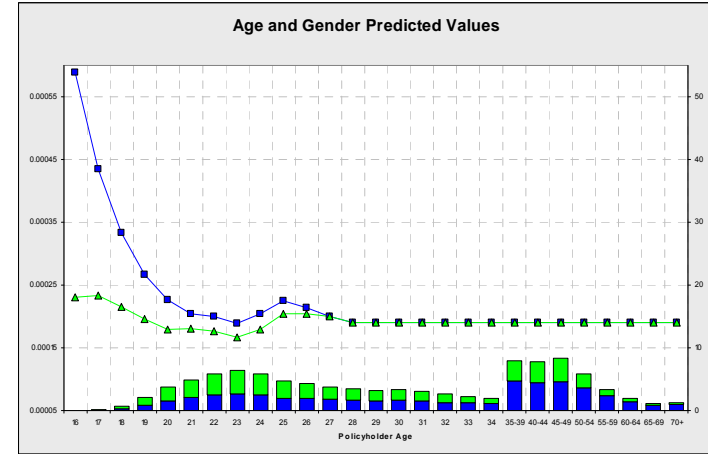
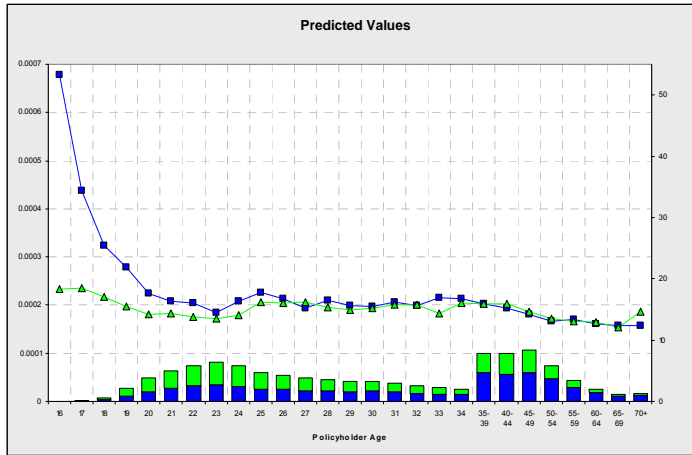


- With the age factor removed, male (the base gender) relatives do not vary by age

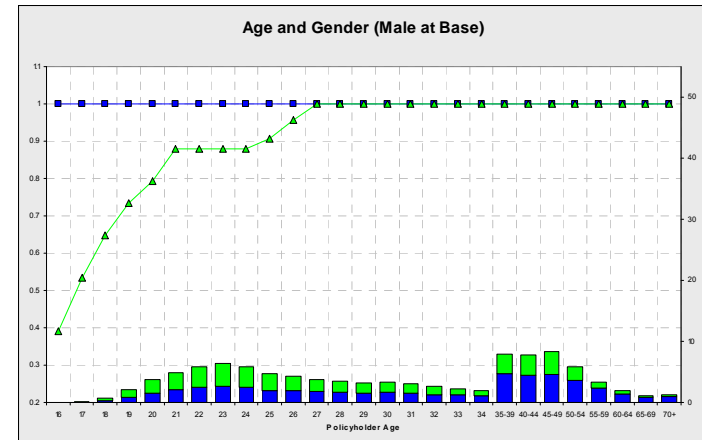
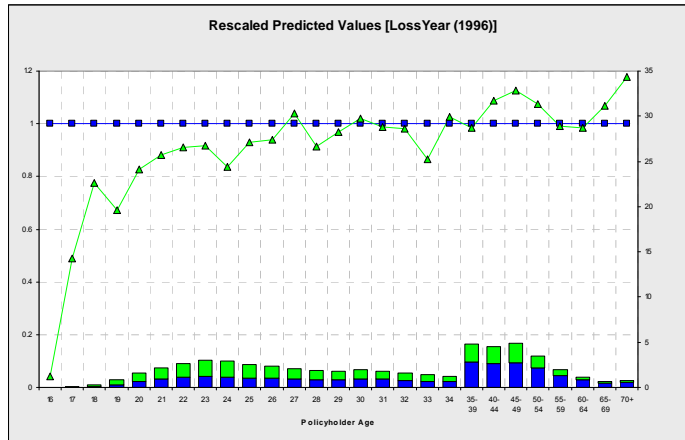
- Because of the interaction term, the female (non-base gender) relatives are unchanged from the full interaction model



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

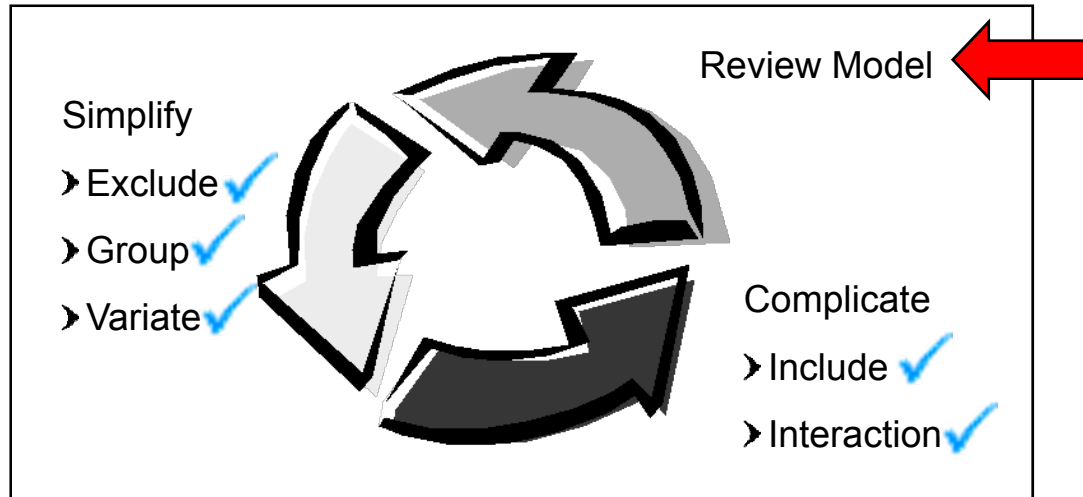


- Simplify the relationship between males and females



# Building the “Best” Model

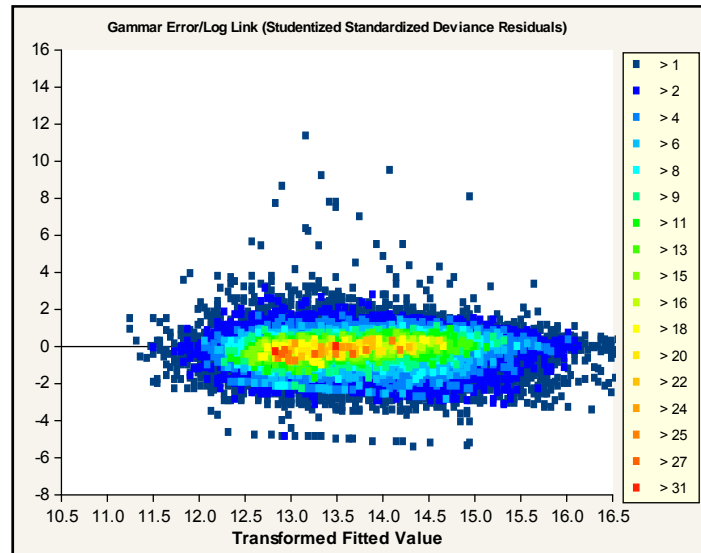
- Modeling is an iterative process



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

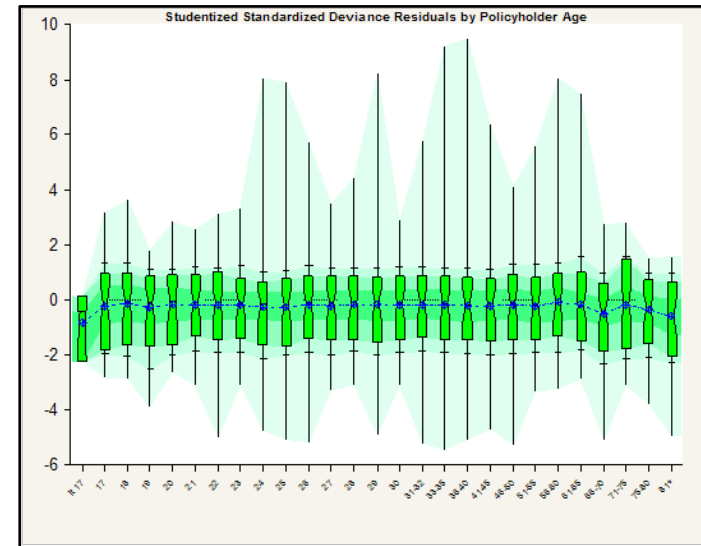
- Tend not to use stepwise regressions for decision-making, but can be useful to identify factors that should be re-tested
  - Backwards: quick review of included factors to see if should consider excluding any
  - Forwards: identifies excluded factors that may be meaningful to include
- Several common statistics can be used for this testing ( $X^2$ , F-test, AIC, etc)

- Re-check residuals to ensure appropriate shape



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

- Does the Box-Whisker show symmetry across levels?

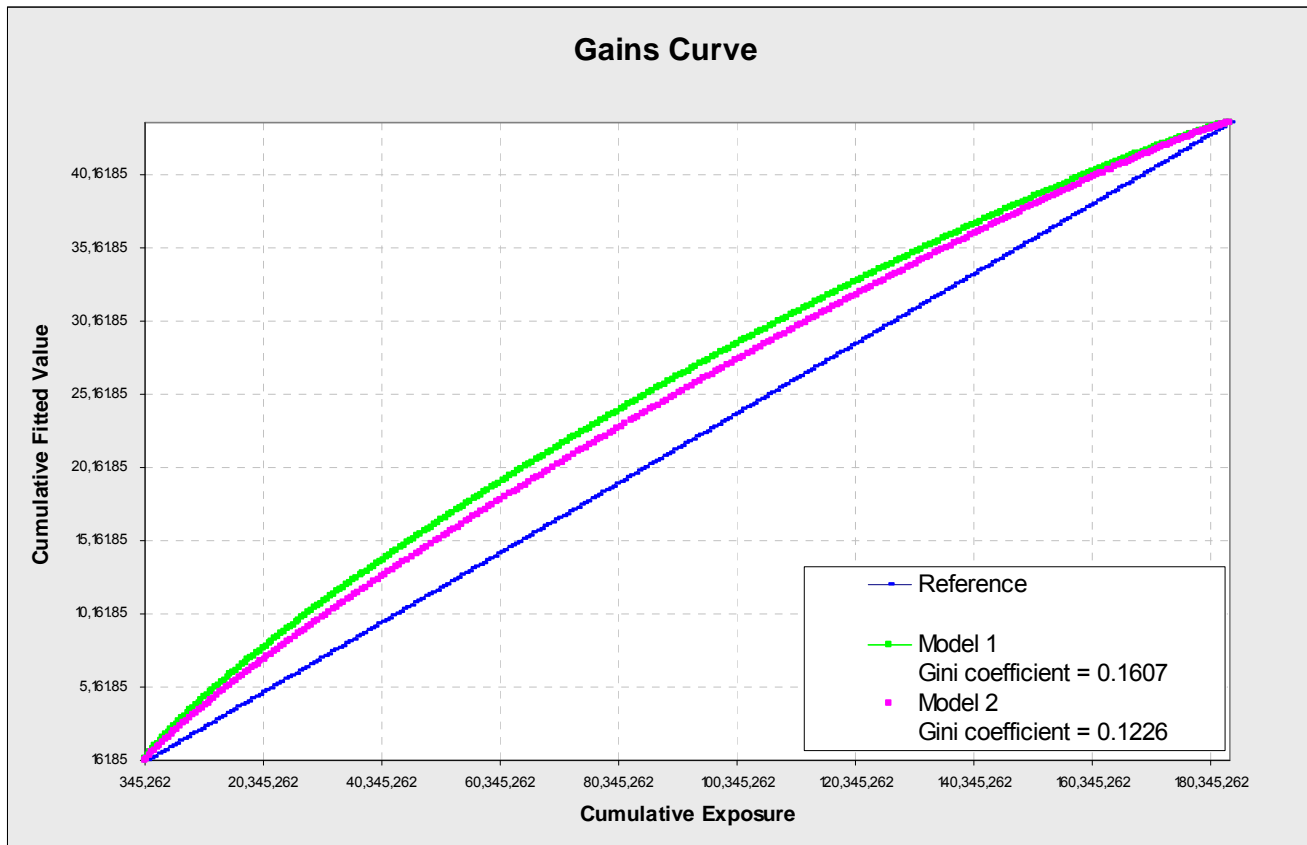


# Validate Model

## Gains Curves



- Compare predictive power of models

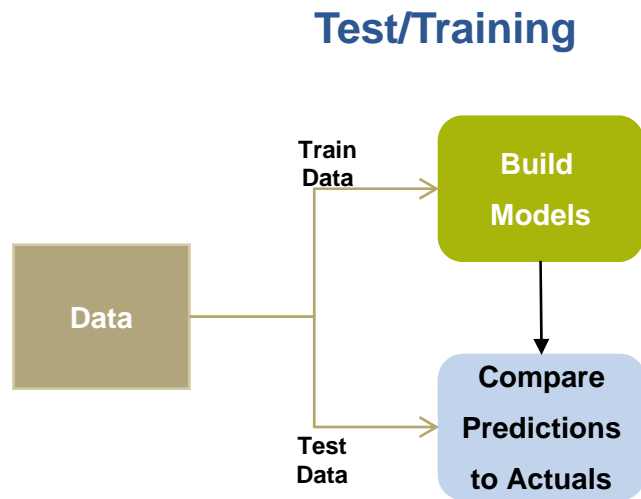


# Validate Model

## Hold-out Samples



- Hold-out samples are effective at validating model
  - Derive parameter estimates based on part of dataset
  - Calculate fitted values on other part of dataset
  - Compare fitted values to historical response

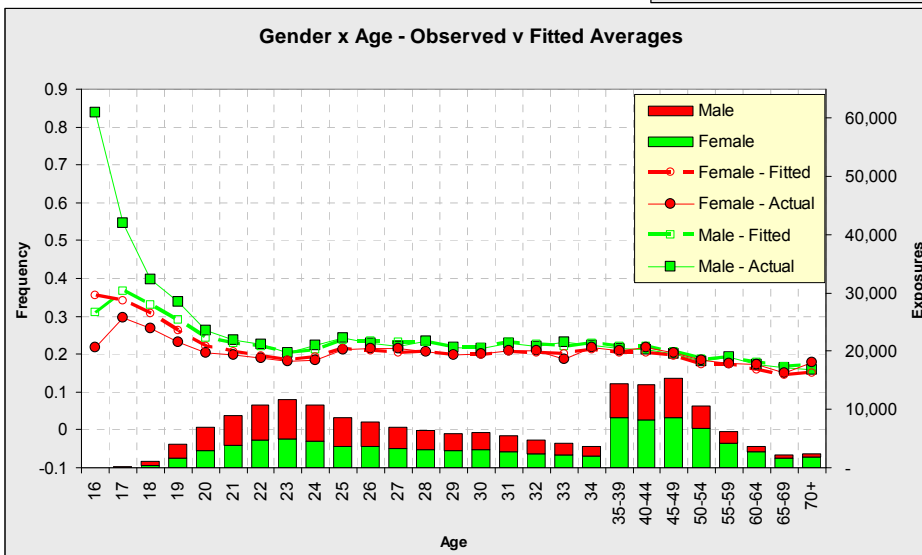
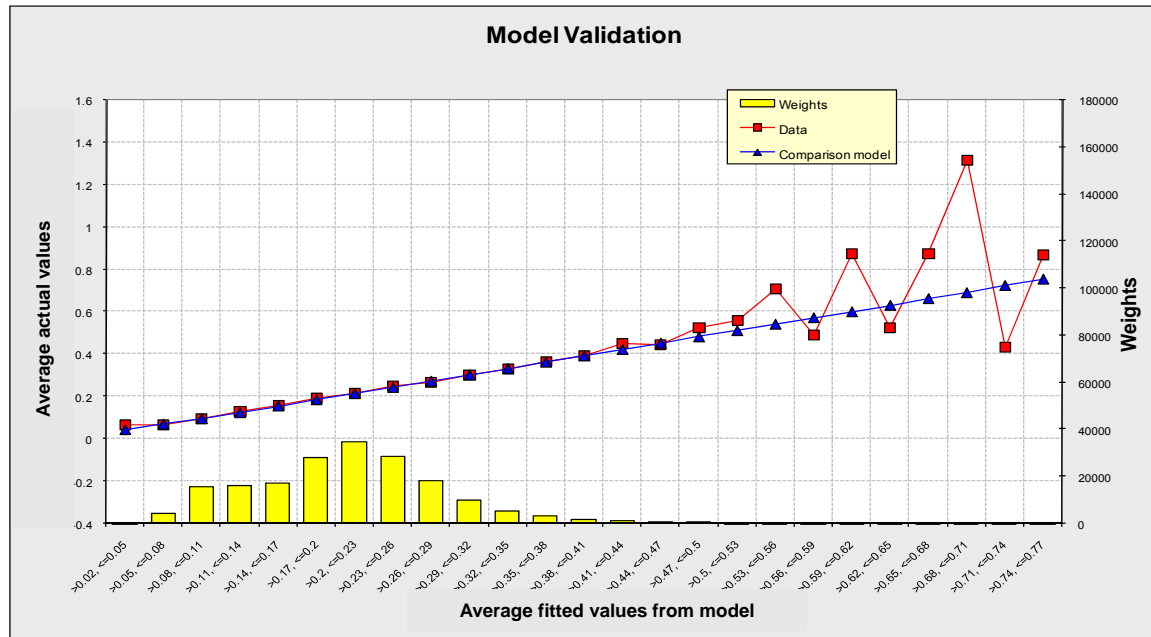


- Larger companies may consider 3 splits
  1. Iterate models
  2. Derive parameters
  3. Validate models/parameters
- Smaller companies may consider a sampling approach

- Predictions should be close to actuals for populated cells

Review of fitted averages can give a hint at areas the model is weak

- View in aggregate (i.e., “lift” chart)



- View by level or by multiple levels

# Combine Predictive Models

## CW Historical Data

Coverage/COL  
Claim Counts  
Exposures  
Characteristics

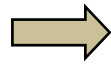
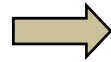
Coverage/COL  
Loss \$ Claim  
Counts  
Characteristics

## CW Predictive Models

Frequency  
Models  
By Coverage/COL

Severity  
Models  
By Coverage/COL

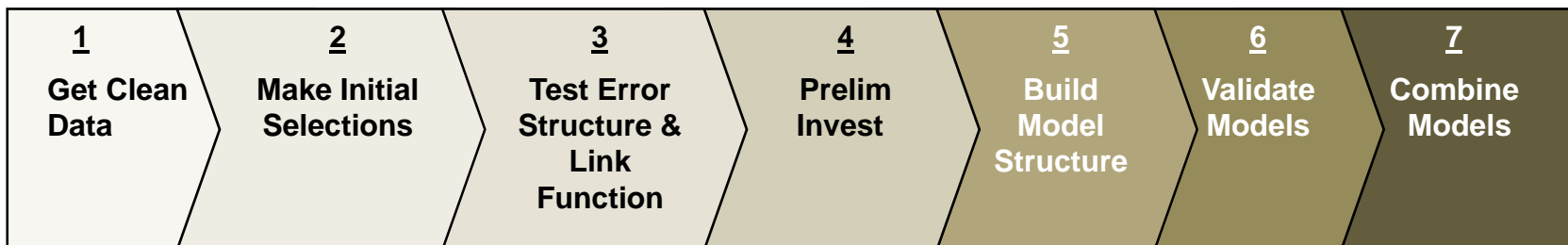
Modeled  
Pure Premiums  
By Coverage/COL



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

# Summary

- GLMs can be a powerful 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



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

---

Give me your name/email after the session

Call me at: (312) 261-9631

Email me at [claudine.modlin@emb.com](mailto:claudine.modlin@emb.com)

### **GLM III will cover:**

- regression splines
- testing the link function
- how to combine GLMs across multiple claim types
- the use of the offset term to constrain models
- techniques for modeling large claims
- practical model validation approaches
- specific issues that arise when modeling price demand elasticity with GLMs, which is of particular importance when undertaking price optimization analyses

## Contact us

---

### **EMB**

12235 El Camino Real  
Suite 150  
San Diego, California  
92130

T +1 (858) 793-1425

F +1 (858) 793-1589

[www.emb.com](http://www.emb.com)



---

© 2008 EMB. All rights reserved. EMB refers to the software and consulting practice carried on by EMB America LLC, EMB Software Management LLP and their directly or indirectly affiliated firms or entities, partnerships or joint ventures, each of which is a separate and distinct legal entity.