EMB America

Dimension Reduction

COSIS Seminar
October 5, 2004
Presented by: Serhat Guven

Agenda

- Background
- Definition
- Rationale
- Techniques
- Conclusion

Special Features of P&C Insurance

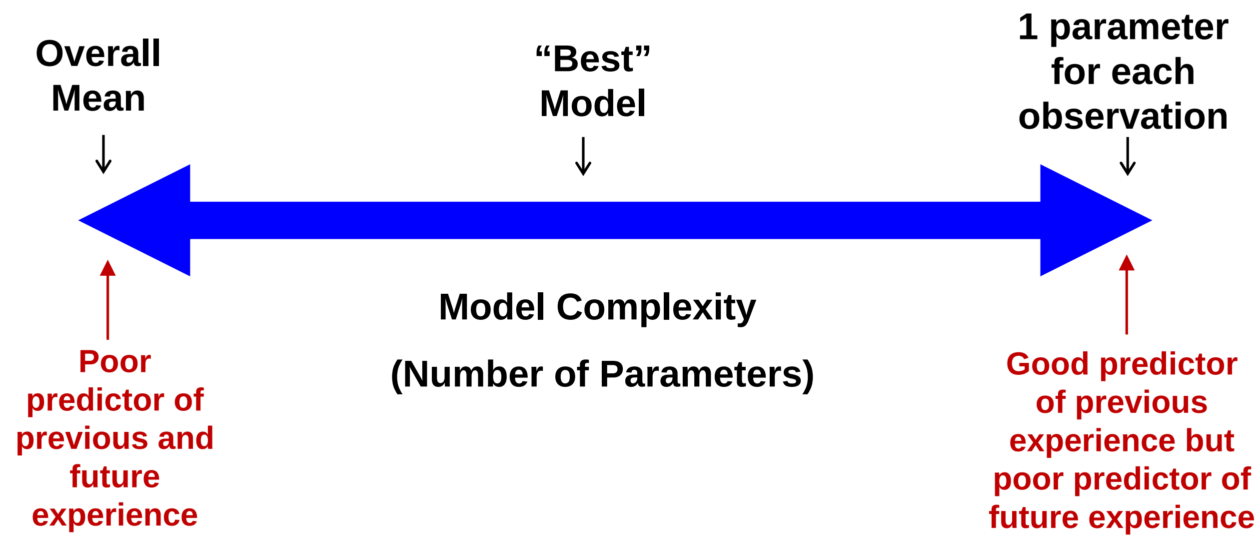
- Low frequency
- Skewed loss distributions
- Often large coefficients of variation
- No natural categories – need continuous estimate of risk rates
- Predictive Models used must recognize these features

Dimension Reduction

- Background
- Definition
- Rationale
- Techniques
- Conclusion

Goal of a Predictive Model

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

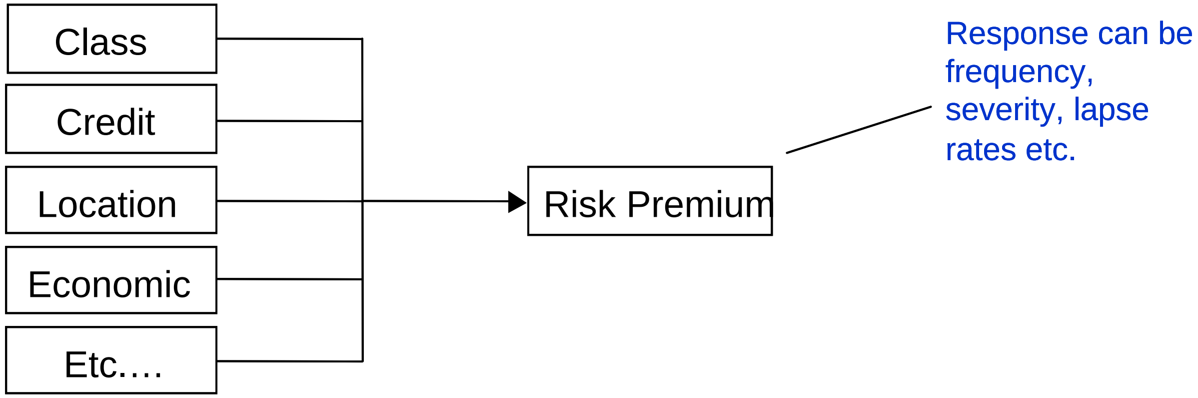


Dimension Reduction

- Background
- Definition
- Rationale
- Techniques
- Conclusion

Goal of a Predictive Model

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



- Larger data storage capabilities allow for a greater number of rating and underwriting variables to be tracked and analyzed.

Dimension Reduction

- Background
- Definition
- Rationale
- Techniques
- Conclusion

Multitude of Factors

- Many factors have been found to be predictive of frequency and/or loss severity. Here are a few for auto…

Safety Features
Limits/Deductibles
Annual Mileage
Other Products
Sales Channel
Coverage
Garaged
Territory
Vehicles
Model Year
Marital Status
Major Convictions
Household Driving Record
Prior Insurance
Traffic Density

Driver Training
Credit Score
Inexperienced Operator
Comp Claims
Occupation
Usage
DWI
Renewals
Good Student
Gender
Not-at-Fault Accidents
Urban v. Rural
Age
Anti-theft Devices
Minor Convictions
Lifestyle Factors
Make/Model
Method of Payment
Homeownership
Drivers
Tier

- Many of these have a significant number of levels.

Multitude of Factors

- Advanced techniques and technology enable the analyst to look at more explanatory variables than previously imagined.
- There still are limitations associated with multivariate approaches.
  - Low volumes of data across dimensions
  - Variables with a large number of rating levels
    - Amount of Insurance
    - Postcode
    - Age
  - Highly collinear variables
- Incorporate Dimension Reduction techniques into the multivariate solution.

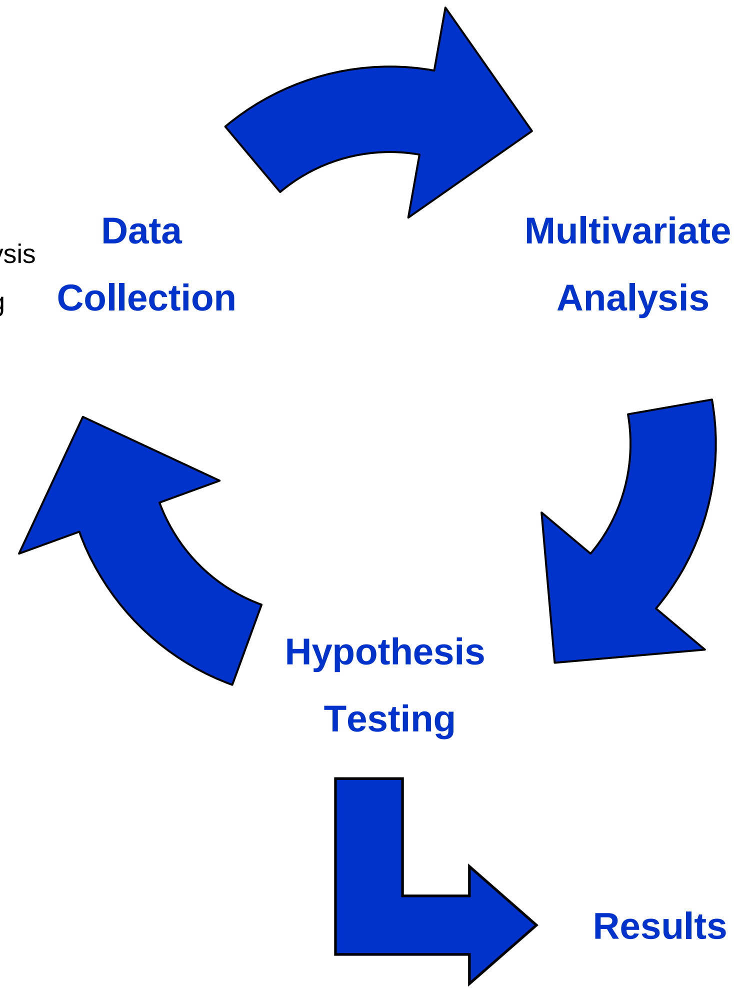
What is Dimension Reduction

- Definition
  - Reducing the dimensionality of a data set by extracting a number of underlying factors, dimensions, clusters, etc., that can account for the variability in the data set.
- Given a table of data:
  - Columns represent both the dimensions and facts of the data.
  - Rows represent the observation.
- Dimension Reduction focuses on reducing both the number of columns (associations among variables) and the number of rows (associations among observations).

What is Dimension Reduction

- Dimension Reduction in the Modeling Process:

>Pre Modeling
- >Principal Components
- >Exploratory Factor Analysis
- >Multidimensional Scaling
- >Clustering



>During Modeling
- >Eliminating
- >Grouping
- >Curve Fitting

Rationale for Dimension Reduction

- Data Storage
  - Advances in warehousing has led to large quantities of data to process.
- Ease of Interpretation
  - Difficulty in Visualizing an n-dimensional rating structure space.
- Collinearity
  - Some degree of redundancy or overlap among rating variables (e.g. Multi Car discounts and # of Vehicles on Policy).
  - Causes a loss in explanatory power.
  - Makes interpretation more difficult.
  - Requires more data to disentangle the individual effects of each variable.

Rationale for Dimension Reduction

- Curse of Dimensionality
  - Cartesian product of the number of rating levels grows exponentially with the inclusion of each rating level.
  - Exposure distribution is not large enough to cover the entire space.

- Principle of Parsimony
  - When two models have the same degree of explanatory power then the simpler model should be selected.

Dimension Reduction Techniques

Association among Variables

- Selection
  - Elimination
  - Grouping
  - Stepwise Regression
    - Backward Elimination
    - Forward Selection
- Transformation
  - Curve Fitting
  - Principle Components
  - Factor Analysis (including Confirmatory Factor Analysis)

Dimension Reduction Techniques

- Association among Observations
  - Multidimensional Scaling
  - Clustering
- Forced Dimension Reduction

Data Description

Type	Descriptions	
Dimensions	Sex	Duration
	Policyholder Age	Garaged
	Rating Area	Installment Indicator
	Vehicle Age	Use
	Vehicle Group	Non standard Indicator
	Driver Restrictions	Major Convictions
	NCD	Minor Convictions
	Protected	MTA Indicator
	Experience	Time
Facts	Exposures	
	Claims	
	Losses	

Dimension Reduction

- Background
- Definition
- Rationale
- Techniques
- Conclusion

Association Among Variables: Selection

Elimination

- Excluding factors entirely is the easiest and most straight-forward way to simplify a model.
- Things to look for:
  - Parameter estimates
    - All parameter estimates are small.
    - All parameter estimates are within two standard errors of zero (i.e., the standard error percentages are all > 50%).
    - Sensible Patterns.
  - Consistency Over Time
  - Models with and without the factor are not significantly different.
    - Chi Square Tests

Dimension Reduction

- Background
- Definition
- Rationale
- Techniques
- Conclusion

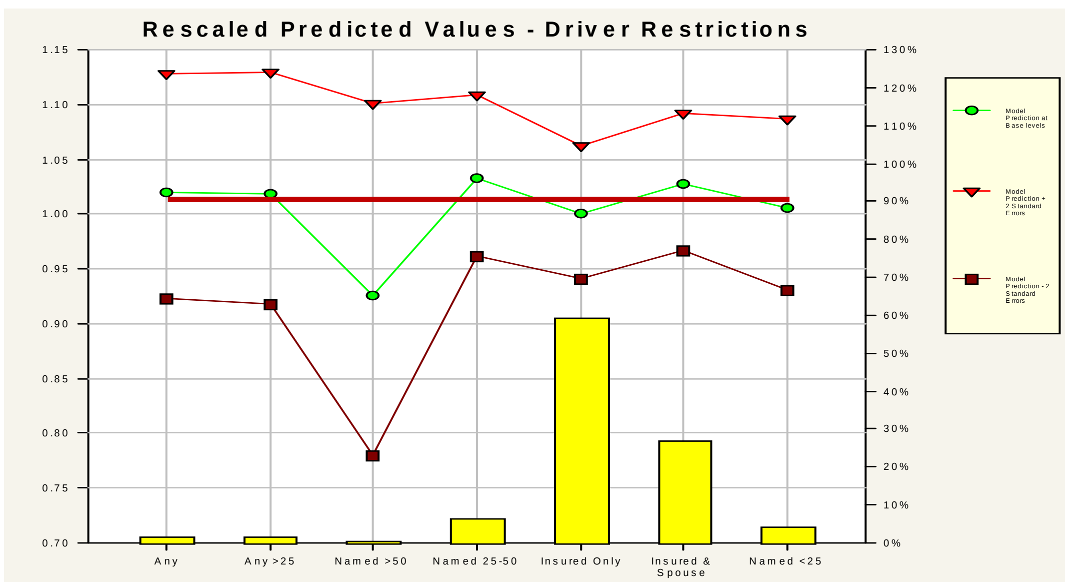
Elimination Example

- Parameter estimates

All close to 0, except Named>50

Name	Value	Standard Error	Standard Error (%)	Weight (%)	Exp(Value)
Driver Restrictions (Any)	0.0198	0.0424	214.3%	1.6%	1.0200
Driver Restrictions (Any >25)	0.0184	0.0440	238.9%	1.4%	1.0186
Driver Restrictions (Named >50)	(0.0768)	0.0816	-106.3%	0.4%	0.9261
Driver Restrictions (Named 25-50)	0.0323	0.0222	68.7%	6.3%	1.0328
Driver Restrictions (Insured Only)				59.3%	
Driver Restrictions (Insured & Spouse)	0.0270	0.0129	47.8%	27.0%	1.0274
Driver Restrictions (Named <25)	0.0056	0.0276	489.4%	4.1%	1.0056

Lowest standard error % is 48%

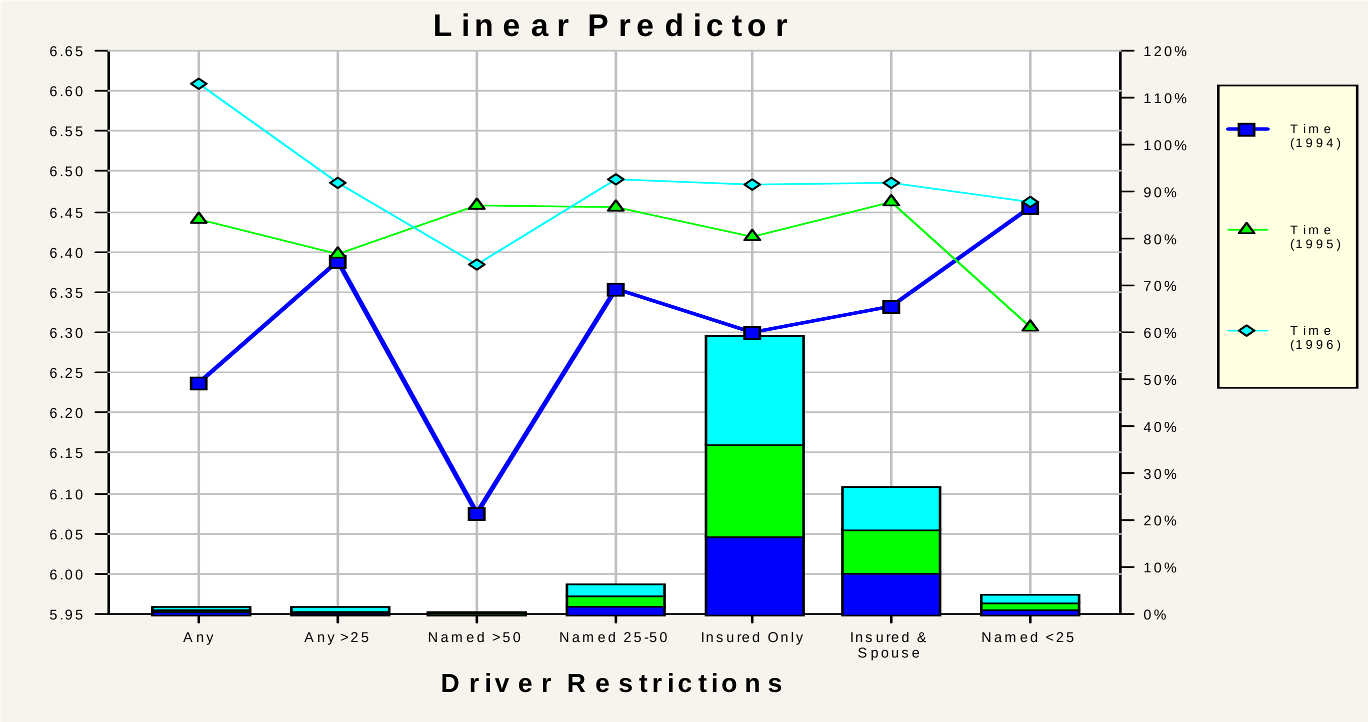


Dimension Reduction

- Background
- Definition
- Rationale
- Techniques
- Conclusion

Elimination Example

- Consistency over time



Elimination Example

- Models with and without the factor are not significantly different.

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%

- Increase in deviance is due to a decrease in the number of parameters.
- H_0: The two models under consideration are not significantly different.

Dimension Reduction

- Background
- Definition
- Rationale
- Techniques
- Conclusion

Association Among Variables: Selection

Grouping

- While a factor might be significant, it may be possible to band certain levels within a factor to create a more parsimonious model.
- Things to look for:
  - Parameter estimates
    - Parameter estimates that are not significantly different from each other.
    - Levels where there is low exposure.
    - Sensible Patterns.
  - Consistency Over Time
  - Models with and without the factor are not significantly different.
    - Chi Square Tests

Dimension Reduction

- Background
- Definition
- Rationale
- Techniques
- Conclusion

Grouping Example

- Standard error of the parameter differences help identify potential groupings

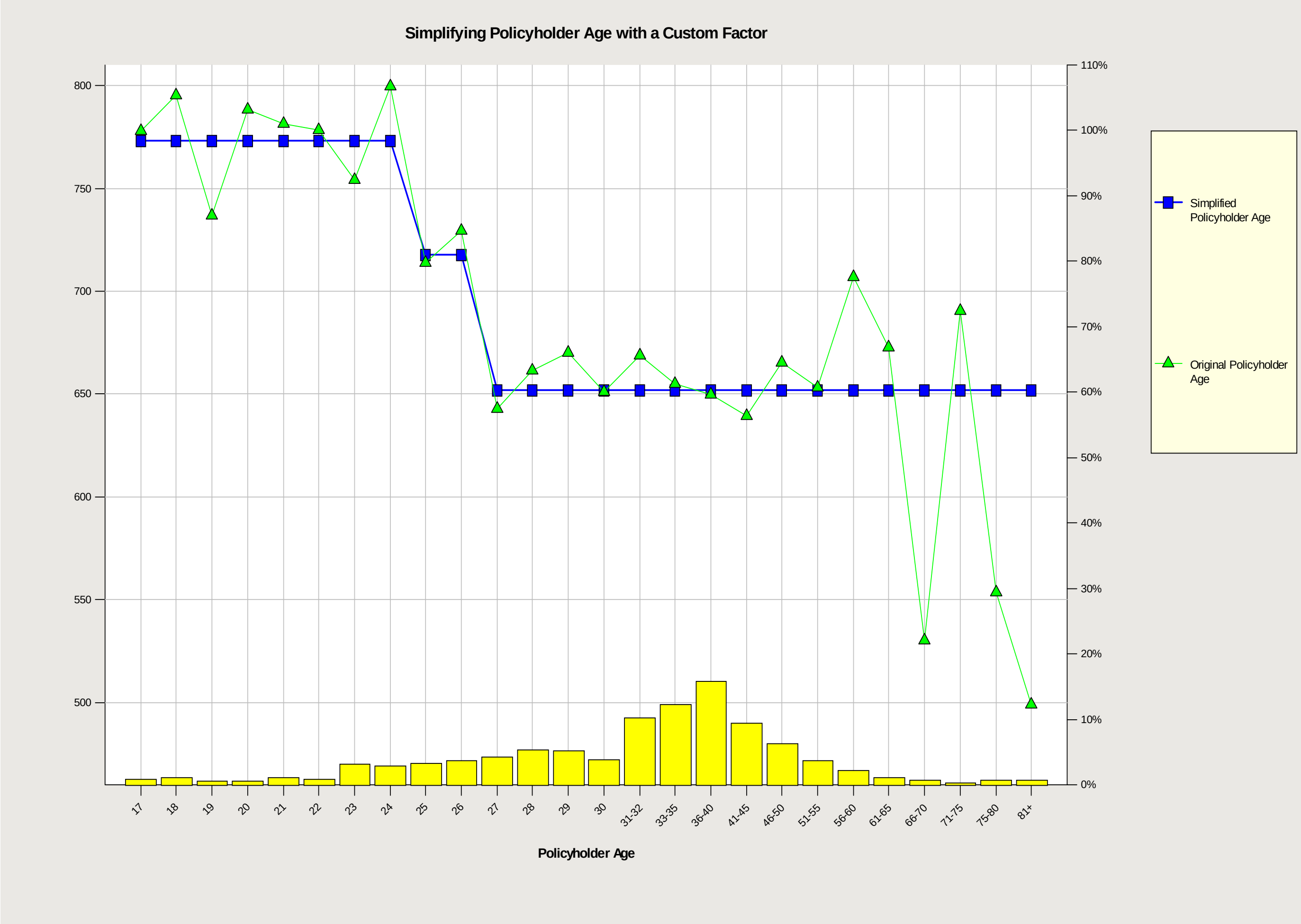
	Policyholder Age (lt 17)	Policyholder Age (17)	Policyholder Age (18)	Policyholder Age (19)	Policyholder Age (20)	Policyholder Age (21)	Policyholder Age (22)	Policyholder Age (23)	Policyholder Age (24)	Policyholder Age (25)	Policyholder Age (26)	Policyholder Age (27)
Policyholder Age (lt 17)												
Policyholder Age (17)	92.3											
Policyholder Age (18)	87.3	308.0										
Policyholder Age (19)	110.0	132.4	91.0									
Policyholder Age (20)	94.1	1,414.6	277.0	154.2								
Policyholder Age (21)	97.7	333.2	153.0	196.1	468.5							
Policyholder Age (22)	97.7	357.7	162.6	200.2	512.2	10,113.7						
Policyholder Age (23)	104.4	130.8	79.6	498.2	158.5	205.1	213.3					
Policyholder Age (24)	90.2	912.9	378.7	101.2	530.3	188.0	204.9	68.4				
Policyholder Age (25)	108.2	104.4	67.8	4,227.8	123.3	141.4	148.0	307.4	56.6			
Policyholder Age (26)	101.6	161.6	90.9	293.4	203.2	322.2	330.4	388.3	82.2	161.7		
Policyholder Age (27)	147.1	41.4	31.8	77.9	46.1	41.5	44.2	35.7	23.5	38.5	29.4	
Policyholder Age (28)	134.7	48.0	35.9	103.7	54.0	49.4	52.7	44.1	26.4	49.4	35.3	132.5
Policyholder Age (29)	129.7	52.4	38.7	123.8	59.1	55.0	58.7	51.2	28.9	59.2	40.6	91.8
Policyholder Age (30)	147.2	41.6	32.0	78.2	46.3	41.8	44.6	36.6	24.1	39.8	30.7	38,134.8
Policyholder Age (31-32)	132.0	48.8	36.0	110.9	55.2	50.3	54.0	43.8	25.6	49.8	34.7	97.8
Policyholder Age (33-35)	142.4	41.8	31.5	82.5	46.9	41.7	44.8	34.2	22.0	37.1	27.6	345.9
Policyholder Age (36-40)	147.6	39.1	29.6	73.9	43.9	38.6	41.5	30.7	20.5	33.0	25.1	2,196.5
Policyholder Age (41-45)	156.5	36.5	28.1	64.9	40.7	35.7	38.3	28.8	19.9	30.5	24.0	192.3
Policyholder Age (46-50)	135.3	47.6	35.7	102.1	53.6	49.2	52.5	44.2	26.6	49.9	36.1	145.0
Policyholder Age (51-55)	147.1	42.0	32.5	79.0	46.8	42.6	45.2	37.9	24.9	41.6	32.3	43,108.9
Policyholder Age (56-60)	114.0	85.6	59.8	481.2	98.7	105.4	110.3	150.3	52.2	254.0	106.9	55.1
Policyholder Age (61-65)	138.1	55.1	43.3	111.9	60.8	59.7	62.1	63.5	38.6	73.0	55.0	288.4
Policyholder Age (66-70)	652.1	23.6	20.7	30.6	25.1	23.2	24.0	21.6	18.3	22.3	20.4	31.9
Policyholder Age (71-75)	127.5	95.4	75.5	243.1	103.9	114.7	116.3	153.2	78.0	194.2	129.6	191.3
Policyholder Age (75-80)	431.4	25.4	22.1	33.8	27.1	25.2	26.0	23.5	19.6	24.4	22.2	36.9
Policyholder Age (81+)	1,822.1	19.7	17.5	24.6	20.8	19.3	19.9	17.8	15.6	18.3	17.0	23.8

Dimension Reduction

- Background
- Definition
- Rationale
- Techniques
- Conclusion

Grouping Example

- Simplify trends in rating factors in order to remove random noise, by grouping factor levels…



Grouping Example

- Models with and without the factor are not significantly different.

Model	Ungrouped	Grouped
Deviance	8,906.4414	8,934.1620
Degrees of Freedom	18,469	18,493
Scale Parameter	0.4822	0.4823
Chi Square Test		27.2%

- Increase in deviance is due to a decrease in the number of parameters.
- H_0: The two models under consideration are not significantly different.

Dimension Reduction

- Background
- Definition
- Rationale
- Techniques
- Conclusion

Association Among Variables: Selection

Stepwise Regression: Backward Elimination

- Build a Model with all variables and delete based on prespecified criteria regarding improvement in model fit:

Model	Variables	Deviance	Degrees of Freedom	Chi Squared Compare to Base
Base	All	8,906.44	18,469	
1	All excl Gender	8,907.09	18,471	65.2%
2	All excl Policyholder Age	8,959.74	18,495	0.1%
3	All excl Rating Area	8,951.61	18,484	0.0%
4	All excl Vehicle Age	10,824.07	18,489	0.0%
	. .			
17	All excl MTA Indicator	8,906.45	18,470	92.2%
18	All excl Time	8,982.06	18,471	0.0%

- Remove factor that performed the worst on the Chi Square test. (MTA Indicator)
- Iterate process with the new base model until no further factors indicated removal.

Dimension Reduction

- Background
- Definition
- Rationale
- Techniques
- Conclusion

Association Among Variables: Selection

Stepwise Regression: Forward Selection

- Build a Model with no factors and add based on prespecified criteria regarding improvement in model fit:

Model	Variables	Deviance	Degrees of Freedom	Chi Squared Compare to Base
Base	Mean	12,380.23	18,596	
1	Mean + Gender	12,377.02	18,594	20.1%
2	Mean + Policyholder Age	12,214.88	18,570	0.0%
3	Mean + Rating Area	12,365.50	18,581	47.1%
4	Mean + Vehicle Age	9,997.75	18,576	0.0%
	.			
	.			
17	Mean + MTA Indicator	12,370.30	18,595	0.2%
18	Mean + Time	12,371.45	18,594	0.1%

- Add the factor that performed the best on the Chi Square test. (Policyholder Age)
- Iterate process with the new base model until no further factors indicated removal.

Association Among Variables: Selection

- Drawbacks to Stepwise Regression:
  - Tendency to Overfit the data.
  - Short cuts the exploratory process through which the researcher gains an intuitive feel for the data.
  - Problems in the presence of collinearity.

Association Among Variables: Transformation

Curve Fitting

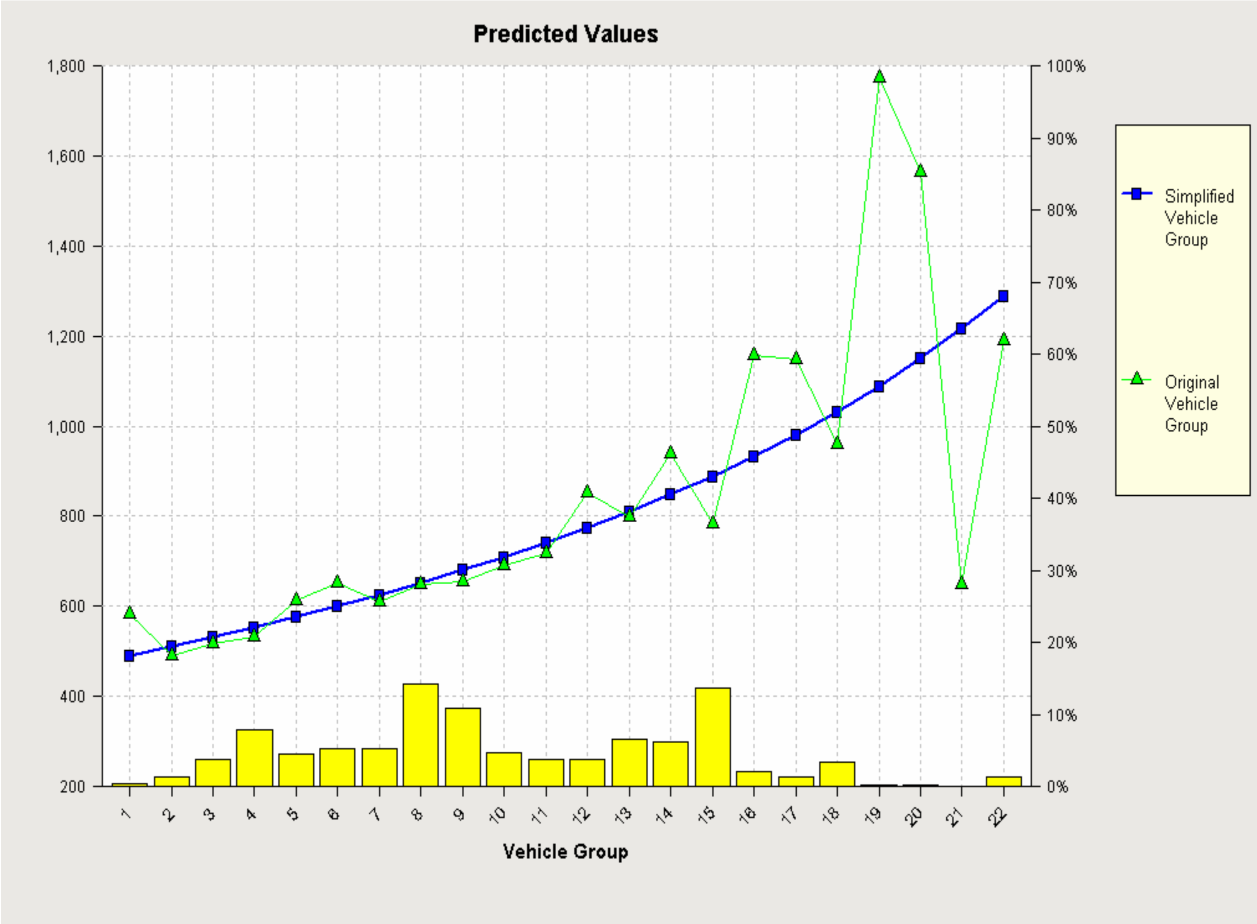
- While a factor might be significant, it may be desirable to smooth adjacent levels to create a more parsimonious model.
- Things to look for:
  - Factors which have a natural x-axis that can be converted to a continuous scale.
  - Factors with a sufficient number of levels to justify curve fitting.
  - Factors with a definite trend or progression.
  - Models with and without the factor are not significantly different.
    - Chi Square Tests

Dimension Reduction

- Background
- Definition
- Rationale
- Techniques
- Conclusion

Modeling-Fitting Curves (Variates)

- Simplify trends in rating factors in order to remove random noise, by fitting an n^{th} degree curve…



Modeling-Fitting Curves (Variates)

- Additional Curve Fitting Options
  - Degree of the Polynomial
  - Multiple curves across the same variable
  - Splines

Dimension Reduction

- Background
- Definition
- Rationale
- Techniques
- Conclusion

Association Among Variables: Transformation

Principle Components Analysis

- Goal: Identify a smaller number of dimensions as a linear combination of the original dimensions that will account for a sufficient amount of information exhibited in the original set.
- Potential Applications
  - Helpful in eliminating collinearity among rating variables
  - Creation of indices from multiple dimensions
  - Identifying patterns of Association among variables
- Linearly combining existing rating factors into a single rating factor.

Dimension Reduction

Principle Components Analysis

- Given an n x p data matrix where n represents the number of observations and p represents the number of rating factors:

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

- Create a standardized matrix X_s such that:

$$x_{s_{i,j}} = \frac{x_{i,j} - \overline{X}_j}{S_j}$$

Principle Components Analysis

- Let Z be an n x c matrix of Principle Components where

$$Z = X_s u$$

  Such that u is a p x c eigenvector matrix.

- The idea is to find u so that Var (Z) is maximized subject to the constraint that $u^T u = 1$
  - This constrained optimization problem is solved with the following equations:

$$Ru = \lambda u$$

  - Where R is the correlation matrix of X_s and λ is the eigenvalue

Principle Components Analysis

- Example: PCA performed on the following factors
  - Vehicle Age (VA)
  - NCD
  - Major Convictions (MJ)
  - Minor Convictions (MN)
- First Principle Component
  - $Z_1 = 0.0687\ VA + -0.7036\ NCD + 0.7038\ MJ + -0.0699\ MN$
  - Z_1 explains about half of the underlying variance in the underlying factors
- Potential Applications
  - Insurance Scores
  - Vehicle Symboling

Association Among Variables: Transformation

Exploratory Factor Analysis

- Goal: Identify underlying source of variance common to two or more variables.
  - Common Factors are unobservable characteristics common to two or more variables.
  - Specific Factors are mutually uncorrelated characteristics specific to only one variable.
- Potential applications
  - Identifying unobservable characteristics.
  - Removing underlying collinearity.
- The idea is to decompose rating variables in linear combinations of latent traits.
  - Factor scores are the location of the original observations in the reduced factor space.

Dimension Reduction

- Background
- Definition
- Rationale
- Techniques
- Conclusion

Exploratory Factor Analysis

- Given the n x p standardized matrix defined earlier then the common factor model is defined as follows:

$$X_s = \Xi \Lambda_c^T + \Delta$$

Such that

$$\Xi = \begin{bmatrix} \xi_1 & \xi_2 & \cdots & \xi_c \end{bmatrix}$$

$$\Delta = \begin{bmatrix} \delta_1 & \delta_2 & \cdots & \delta_p \end{bmatrix}$$

$$\Lambda_c = \begin{bmatrix} \lambda_{1,1} & \lambda_{1,2} & \cdots & \lambda_{1,c} \\ \lambda_{2,1} & \lambda_{2,2} & \cdots & \lambda_{2,c} \\ \vdots & \vdots & & \vdots \\ \lambda_{p,1} & \lambda_{p,2} & \cdots & \lambda_{p,c} \end{bmatrix}$$

Where ξ_j is the j^{th} factor that is common to all observed variables, $\lambda_{i,j}$ is the coefficient and δ_i is the i^{th} factor specific to the i^{th} rating variable.

Exploratory Factor Analysis

- Determine the factor scores for use in the larger multidimensional model:

$$\Xi = X_s R^{-1} \Lambda_c$$

Where R is the correlation matrix of X_s

- Comparing to Principal Components:
  - Principal components assumes that all the variability should be used in the resulting analysis
  - Exploratory Factor analysis assumes that only the variability associated with the common factors should be used in the resulting analysis

Dimension Reduction

- Background
- Definition
- Rationale
- Techniques
- Conclusion

Exploratory Factor Analysis

Potential Applications

- Generating new rating variables
- Simplifying existing rating structures

Association Among Observations

Multidimensional Scaling

- Goal: Detect meaningful underlying dimensions that allow one to explain observed similarities between objects.
- Approach is to arrange objects in a space with a particular number of dimensions so as to produce the observed distances.
- Types
  - Metric
  - Nonmetric
  - Multidimensional Analysis of Preference
- Potential Applications: Perceptual Mappings
  - Identify and model customers premium expectations.
  - Map the importance and influence of various insurance operations based on customer surveys.

Association Among Observations

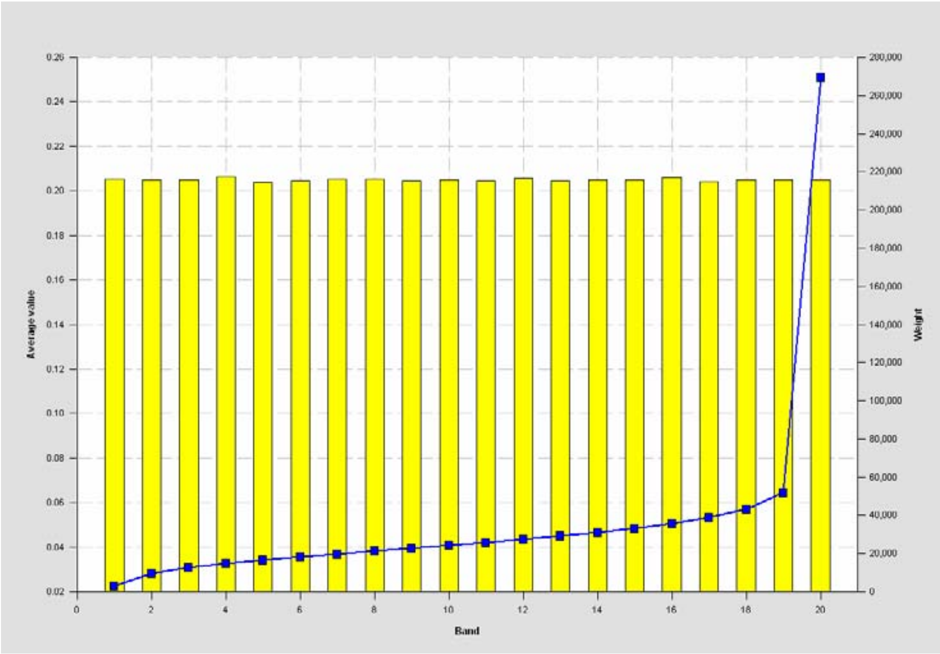
Clustering

- Goal:
  - Minimize within-group heterogeneity.
  - Maximize cross-group heterogeneity.
  - Produce groupings which are predictive in future.
- Basic Methods
  - Quantiles
  - Equal Weight
  - Similarity Methods
  - K-means Clustering

Dimension Reduction

Clustering

- Quantiles
  - Create groups with equal numbers of observations.
- Equal Weight
  - Create groups which have an equal amount of weight.



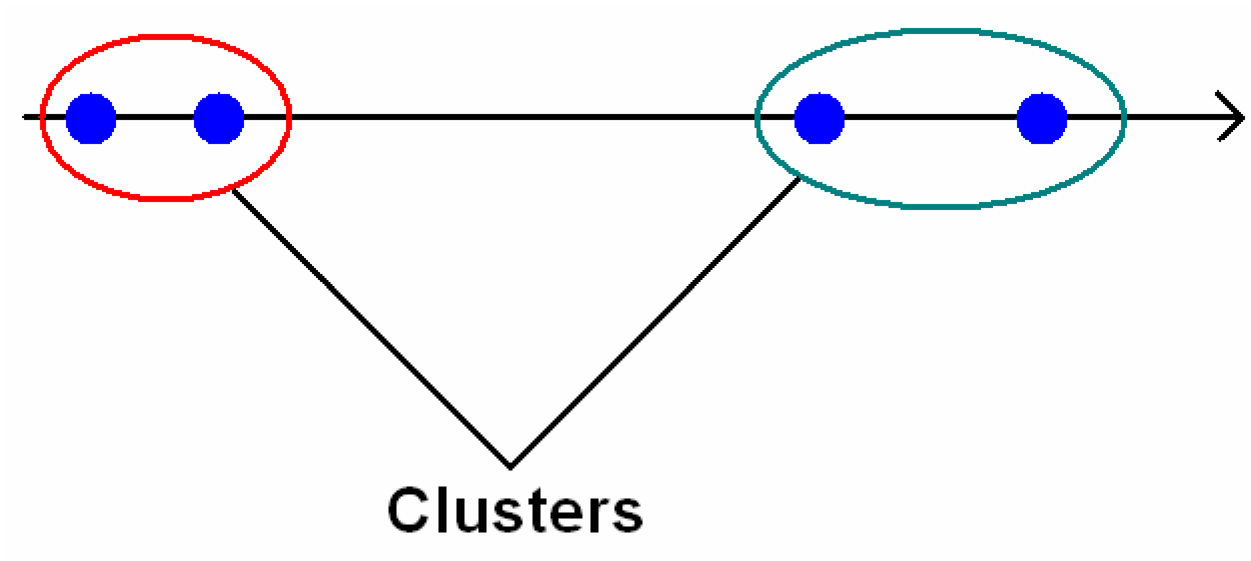
Dimension Reduction

- Background
- Definition
- Rationale
- Techniques
- Conclusion

Clustering

Similarity Methods

- General Approach
  - Rank the data set by the statistic you wish to cluster.
  - Decide on which pair of records are the ‘most similar.’
  - Group these records.
  - Repeat until left with the desired number of groups.



Dimension Reduction

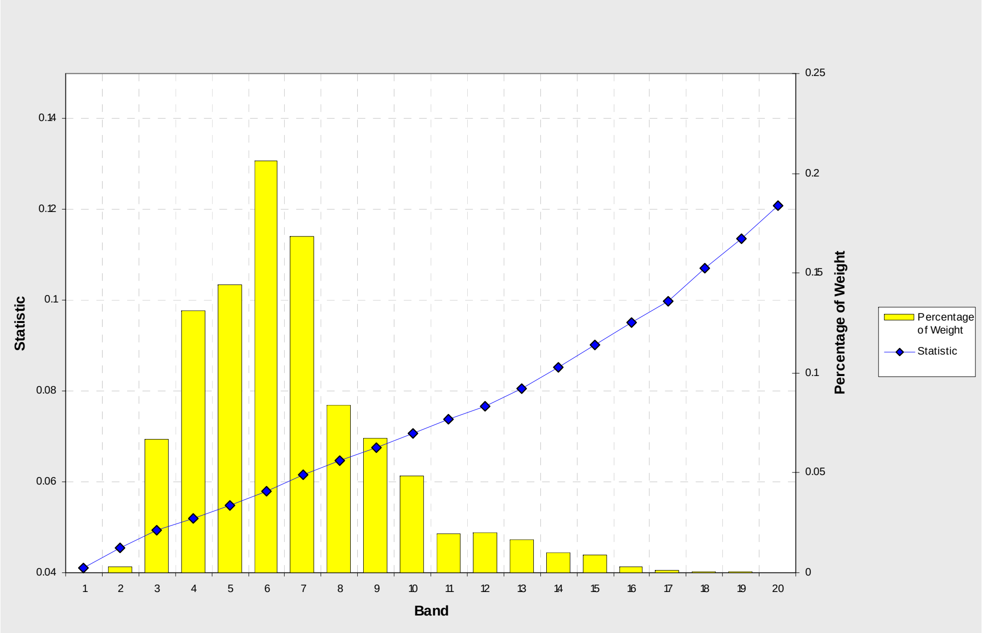
- Background
- Definition
- Rationale
- Techniques
- Conclusion

Clustering

Similarity Methods

Average Linkage

- Distance between clusters is the average distance between pairs of observations, one in each cluster.
- Tends to join clusters with small variances.



Dimension Reduction

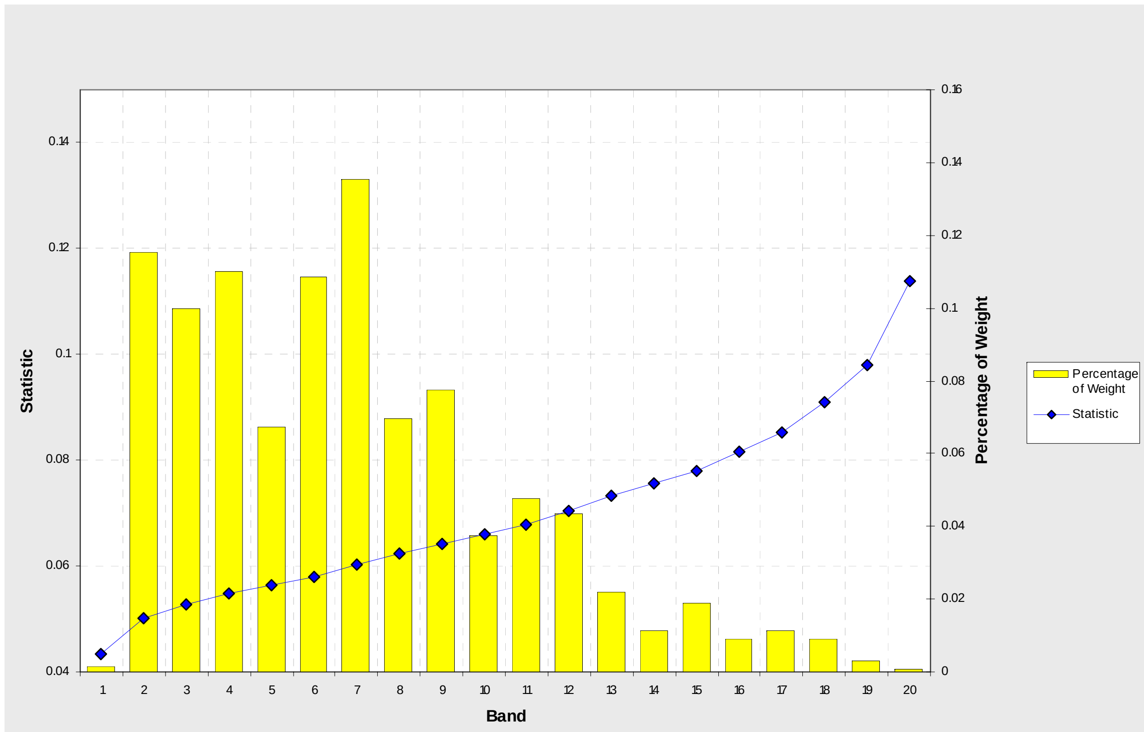
- Background
- Definition
- Rationale
- Techniques
- Conclusion

Clustering

Similarity Methods

Centroid

- Distance between clusters is the difference between the mean values of the clusters squared.



Clustering

- K-means
  - Rank the observations.
  - Split into k groups e.g. using quantile method.
  - Calculate the mean value of each group.
  - Define group start/end-points as being half-way between adjacent mean values.
  - Reallocate each observation.
  - Repeat until group start and end-points converge.

Forced Dimension Reduction

- Regulatory disallows credit {undesirable subsidy}
  - Include in modeling of frequency and severity to get most predictive pure premium
  - Model rating algorithm without credit variable
  - Try to adjust for lack of credit
- Business dimension {desirable subsidy}
  - Model rating algorithm without adjustments as if factor fully included
  - Otherwise, model will try to “correct” for excluded variable

Dimension Reduction

Conclusion

- How many variables are available?
  - Rating plan vs. available in warehouse.
  - Credit factors.
  - Socio demographic.
- Objective is to identify factors which are predictive
  - Which are best at differentiating risks?
  - Understand all predictive variables before building in any constraints.

Conclusion

- How many levels do we have in our predictive variables?
  - Driver age.
  - Zipcode.
  - Numbers of levels and nature of variable will determine most appropriate measure.
- Objective is to identify underlying signal and represent it in our models.

Conclusion

- Available factors.
- Identify predictive variables.
- Extract signal from predictive variables.
- Use models to build rating plan.

Questions?