



Insurance and Actuarial
Advisory Services

 **ERNST & YOUNG**

Quality In Everything We Do

Practical Issues in Model Design

Chuck Boucek

**CAS Seminar on Predictive Modeling – Las Vegas, Nevada
October 11 – 12, 2007**

www.ey.com/us/actuarial

Overview

- Data usually does not seamlessly fit into model assumptions
- The focus of this presentation is the impact that selected issues have on the design matrix
- Agenda
 - Design matrix overview
 - Nonlinearity in predictors
 - Missing data

Design Matrix Overview

Design Matrix
Nonlinearity
Missing Data

- Representation of the predictor variables used to construct model

Data

<u>Class</u>	<u>State</u>	<u>AOI</u>	<u>Pop Density</u>
65198	MA	125	.033
65198	IL	235	.032
70446	MA	240	.034
70446	FL	350	.044
64446	MA	100	.023
64446	IN	110	.025

Design Matrix

<u>Intercept</u>	<u>Class</u>	<u>ST MA</u>	<u>AOI</u>	<u>Pop Density</u>
1	0	0	125	.033
1	0	0	235	.032
1	1	0	240	.034
1	1	0	350	.044
1	0	1	100	.023
1	0	1	110	.025

Source of all graphs: Ernst & Young
Insurance and Actuarial Advisory Services

How is GLM Fit to Data?

Design Matrix
Nonlinearity
Missing Data

Design Matrix \times **Coefficients** = **Linear Predictors**

$$\begin{bmatrix} 1 & 0 & 0 & 1 & 125 & .033 \\ 1 & 0 & 0 & 0 & 235 & .032 \\ 1 & 1 & 0 & 1 & 240 & .034 \\ 1 & 1 & 0 & 0 & 350 & .044 \\ 1 & 0 & 1 & 1 & 100 & .023 \\ 1 & 0 & 1 & 0 & 110 & .025 \end{bmatrix} \times \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \\ a_5 \\ a_6 \end{bmatrix} = \begin{bmatrix} LP_1 \\ LP_2 \\ LP_3 \\ LP_4 \\ LP_5 \\ LP_6 \end{bmatrix}$$

- Linear predictors are transformed to estimate response data via inverse link function
- Family and link function determine form of likelihood function (L)

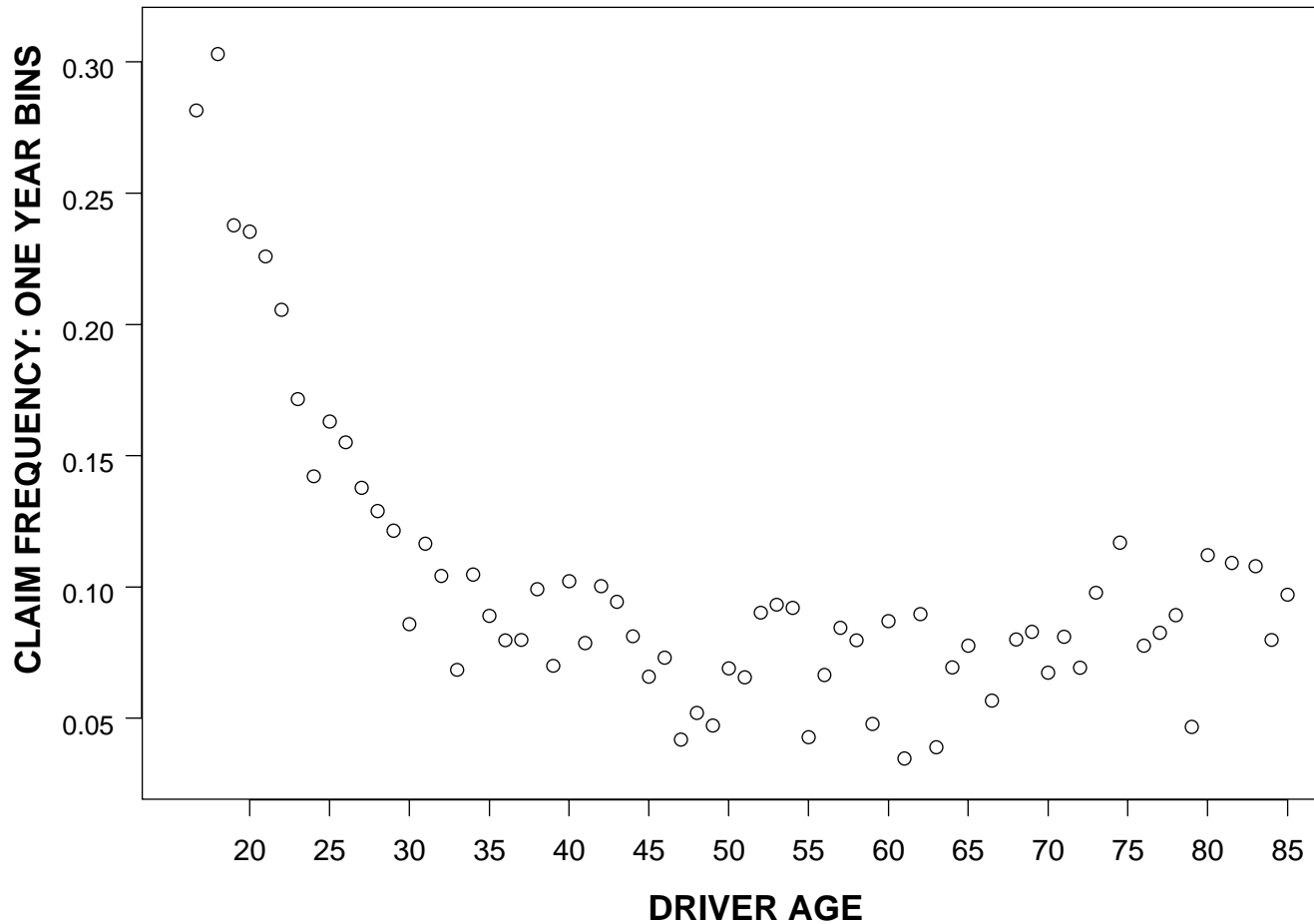
– Family: Gaussian; Link: identity, $-\log(L)$:

$$-l(y) = \sum_{i=1}^n \left(\frac{1}{2} \frac{(y_i - \sum_{j=1}^p x_{ij} a_j)^2}{\sigma^2} + \frac{1}{2} \ln(2\pi\sigma^2) \right)$$

Nonlinearity – Description of Issue

Design Matrix
Nonlinearity
Missing Data

CLAIM FREQUENCY BY DRIVER AGE
MALE PRINCIPAL OPERATOR

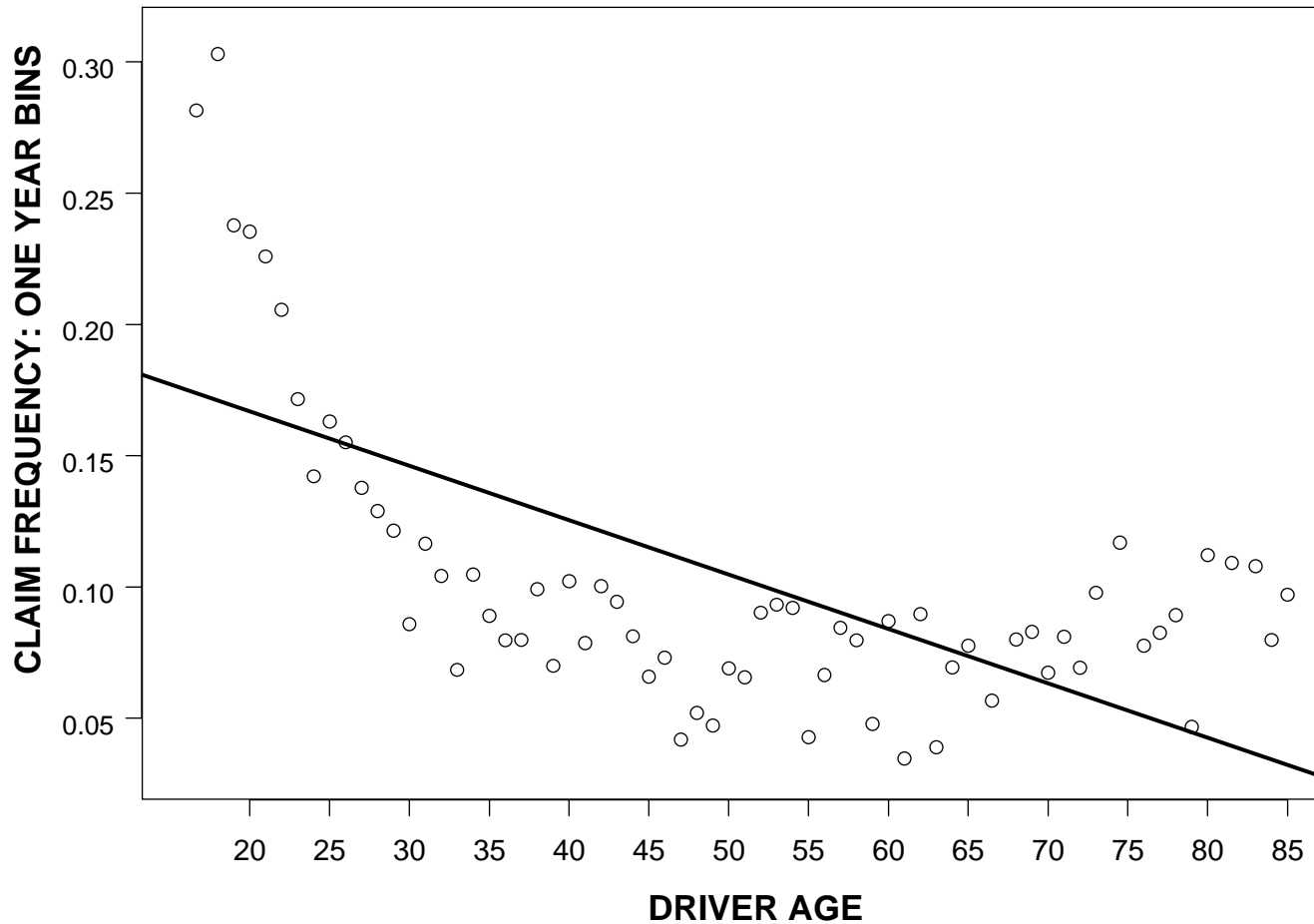


(Continued)

Nonlinearity – Description of Issue

Design Matrix
Nonlinearity
Missing Data

PREDICTED CLAIM FREQUENCY - LINEAR
MALE PRINCIPAL OPERATOR

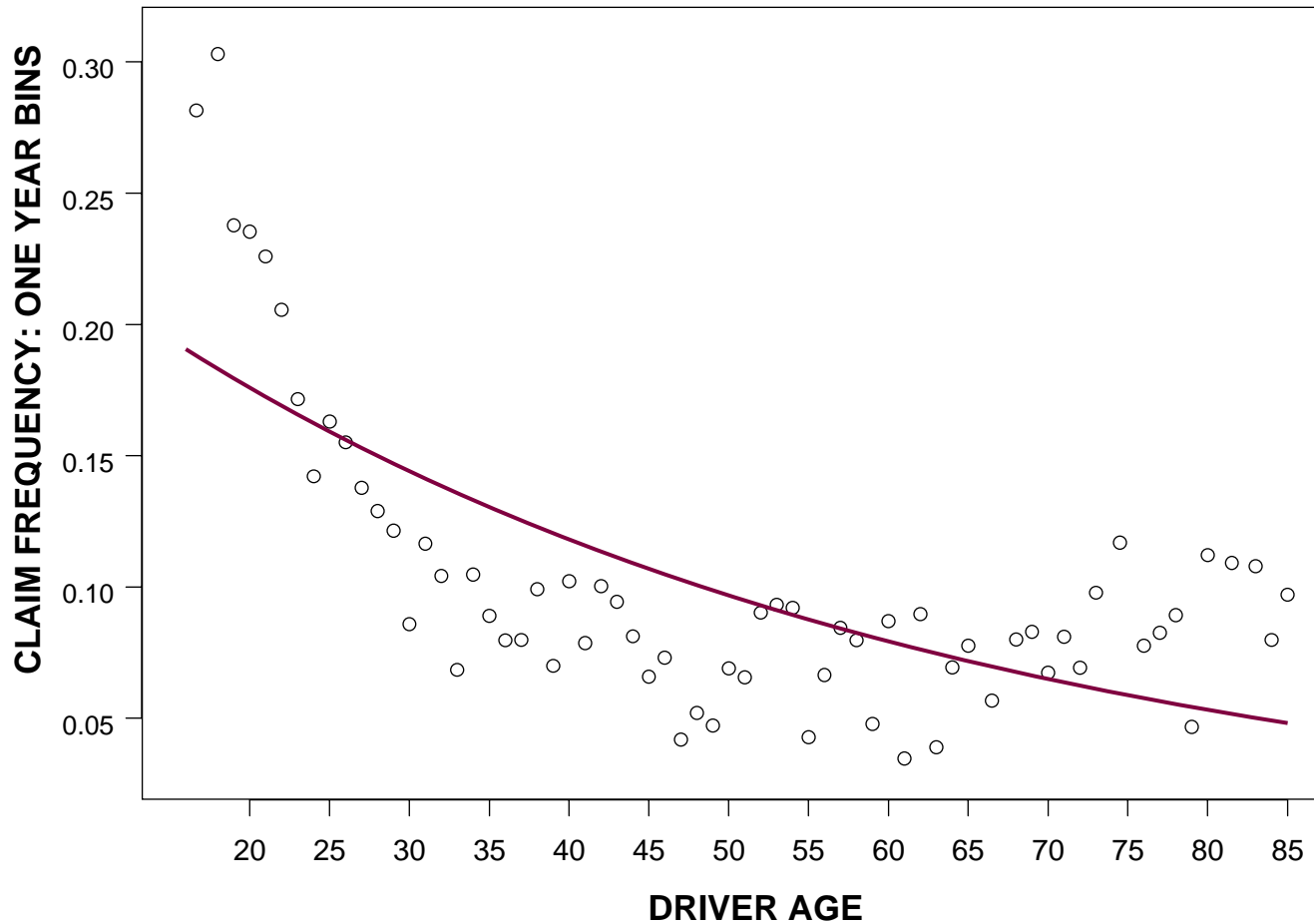


(Continued)

Nonlinearity – Description of Issue

Design Matrix
Nonlinearity
Missing Data

PREDICTED CLAIM FREQUENCY - POISSON GLM
MALE PRINCIPAL OPERATOR



Design Matrix

Design Matrix
Nonlinearity
Missing Data

$$\begin{array}{c} \text{Intercept} \\ \text{Age} \end{array} \begin{bmatrix} 1 & 81 \\ 1 & 17 \\ 1 & 24 \\ 1 & 18 \\ 1 & 83 \\ 1 & 55 \end{bmatrix} \times \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} LP_1 \\ LP_2 \\ LP_3 \\ LP_4 \\ LP_5 \\ LP_6 \end{bmatrix}$$

- One column is added to the design matrix
 - Column represents driver age
- GLM is fit with likelihood and link functions

Nonlinearity – Description of Issue

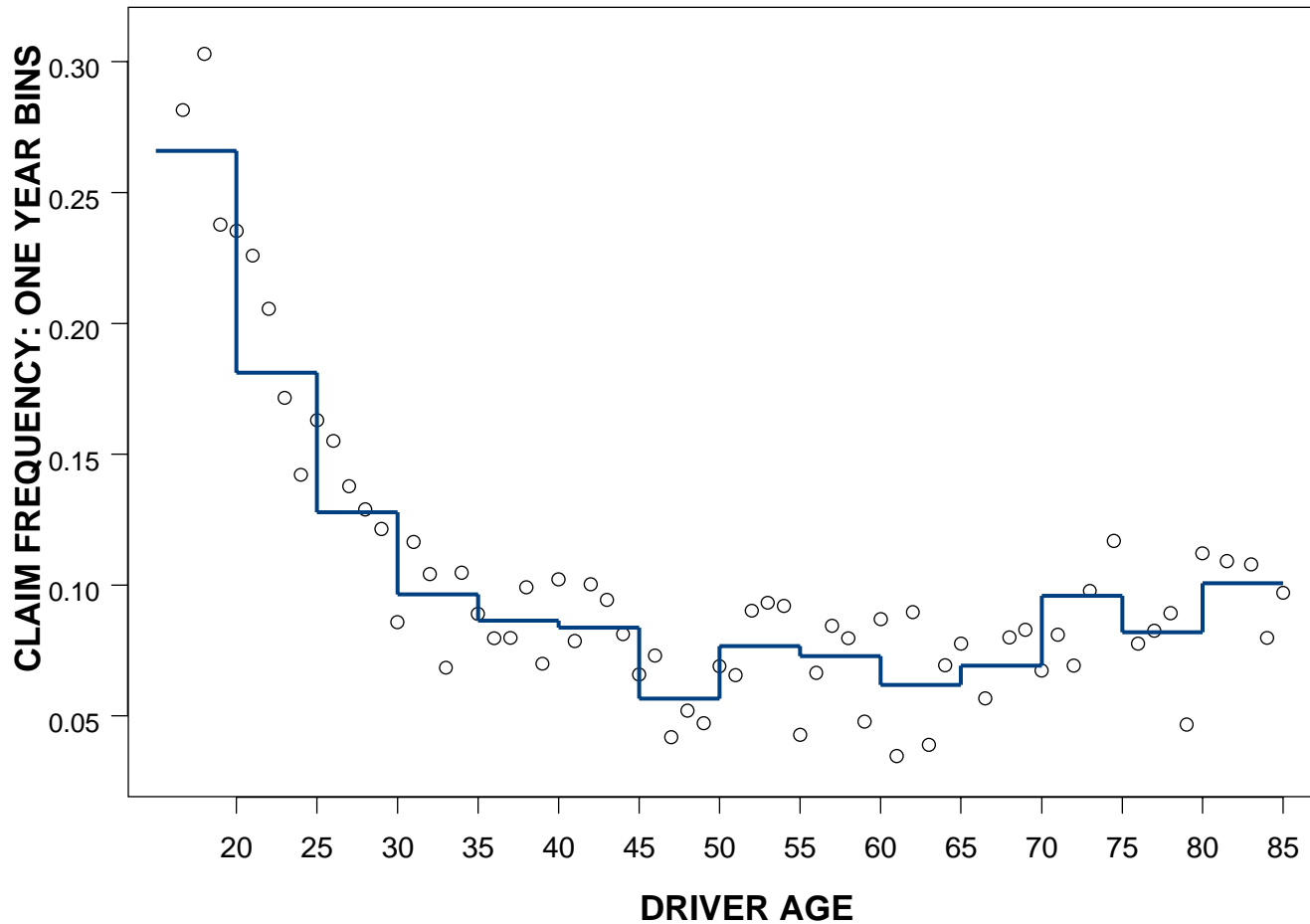
Design Matrix
Nonlinearity
Missing Data

- Three approaches to address nonlinearity
 - Creation of categories (Binning)
 - Polynomial
 - Spline

Nonlinearity – Binning

Design Matrix
Nonlinearity
Missing Data

PREDICTED CLAIM FREQUENCY - BINNING
MALE PRINCIPAL OPERATOR



(Continued)

Design Matrix
Nonlinearity
 Missing Data

Design Matrix with Binning

$$\begin{matrix}
 \text{Intercept (15-20)} & \text{(21-24)} & \dots & \text{(80-85)} \\
 \left[\begin{array}{c|ccc|c}
 1 & 0 & 0 & \dots & 1 \\
 1 & 1 & 0 & \dots & 0 \\
 1 & 0 & 1 & \dots & 0 \\
 1 & 1 & 0 & \dots & 0 \\
 1 & 0 & 0 & \dots & 1 \\
 1 & 0 & 0 & \dots & 0
 \end{array} \right]
 \end{matrix}
 \times
 \begin{bmatrix}
 a_1 \\
 a_2 \\
 a_3 \\
 \vdots \\
 \vdots \\
 \vdots \\
 a_{14}
 \end{bmatrix}
 =
 \begin{bmatrix}
 LP_1 \\
 LP_2 \\
 LP_3 \\
 LP_4 \\
 LP_5 \\
 LP_6
 \end{bmatrix}$$

- Thirteen columns are added to the design matrix
 - Each column represents a driver age bin
- GLM is fit with same likelihood and link functions

(Continued)

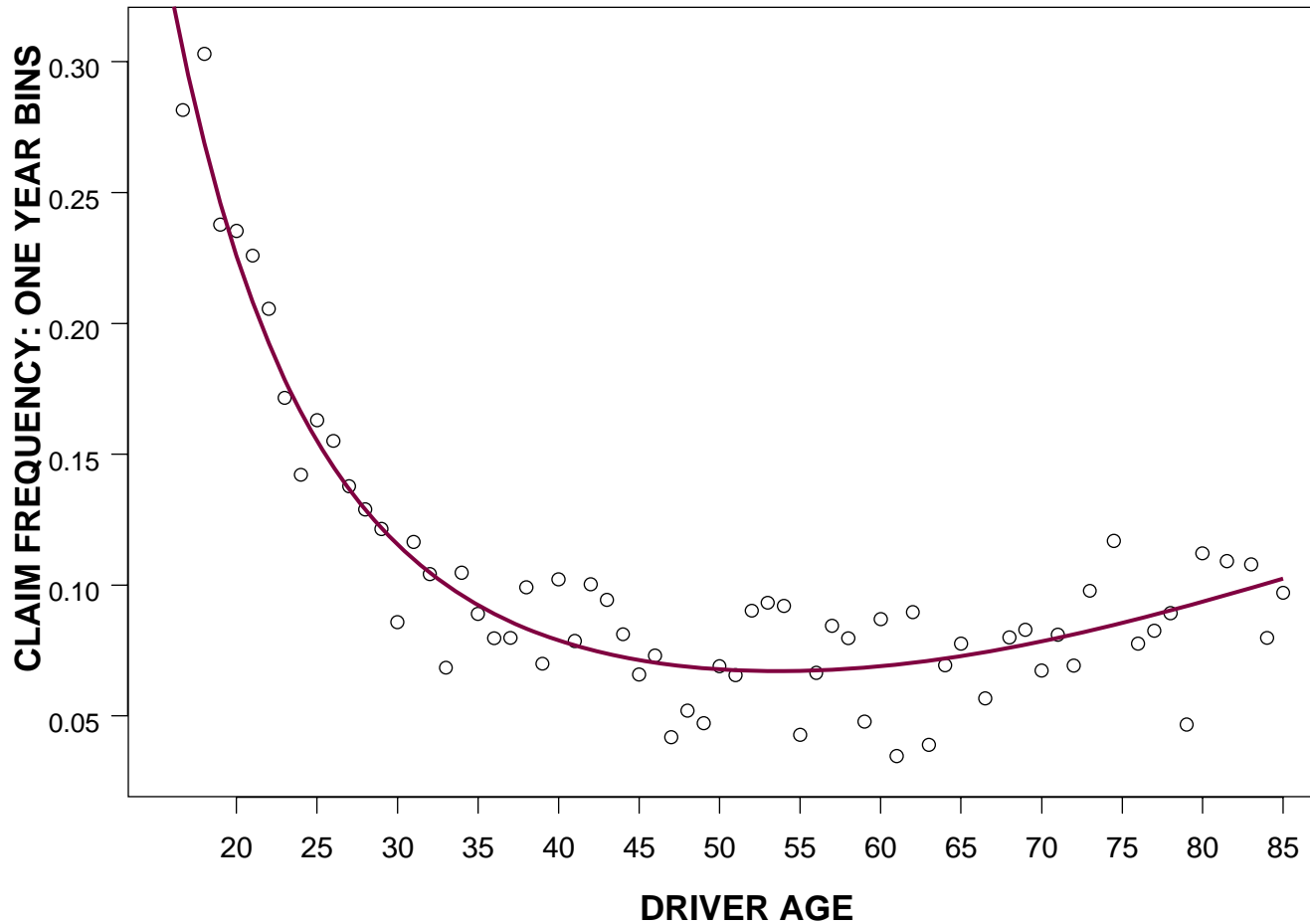
Nonlinearity – Binning

- Primary advantage
 - Simple conceptually
- Primary disadvantages
 - Adds complexity to the model (high # of parameters)
 - Can produce noisy predictions

Nonlinearity – Polynomial

Design Matrix
Nonlinearity
Missing Data

PREDICTED CLAIM FREQUENCY - POLYNOMIAL
MALE PRINCIPAL OPERATOR



(Continued)

Design Matrix with Polynomial

Design Matrix
Nonlinearity
Missing Data

$$\begin{array}{c|ccc} \text{Intercept} & \text{Age} & \text{Age}^2 & \text{Age}^3 \\ \hline \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} & \begin{bmatrix} 81 \\ 17 \\ 24 \\ 18 \\ 83 \\ 55 \end{bmatrix} & \begin{bmatrix} 81^2 \\ 17^2 \\ 24^2 \\ 18^2 \\ 83^2 \\ 55^2 \end{bmatrix} & \begin{bmatrix} 81^3 \\ 17^3 \\ 24^3 \\ 18^3 \\ 83^3 \\ 55^3 \end{bmatrix} \end{array} \mathbf{X} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \end{bmatrix} = \begin{bmatrix} \text{LP}_1 \\ \text{LP}_2 \\ \text{LP}_3 \\ \text{LP}_4 \\ \text{LP}_5 \\ \text{LP}_6 \end{bmatrix}$$

- Three columns are added to the design matrix
 - Each column represents driver age raised to a power
- GLM is fit with same likelihood and link functions
- An orthogonal polynomial is generally used rather than the above simple polynomial

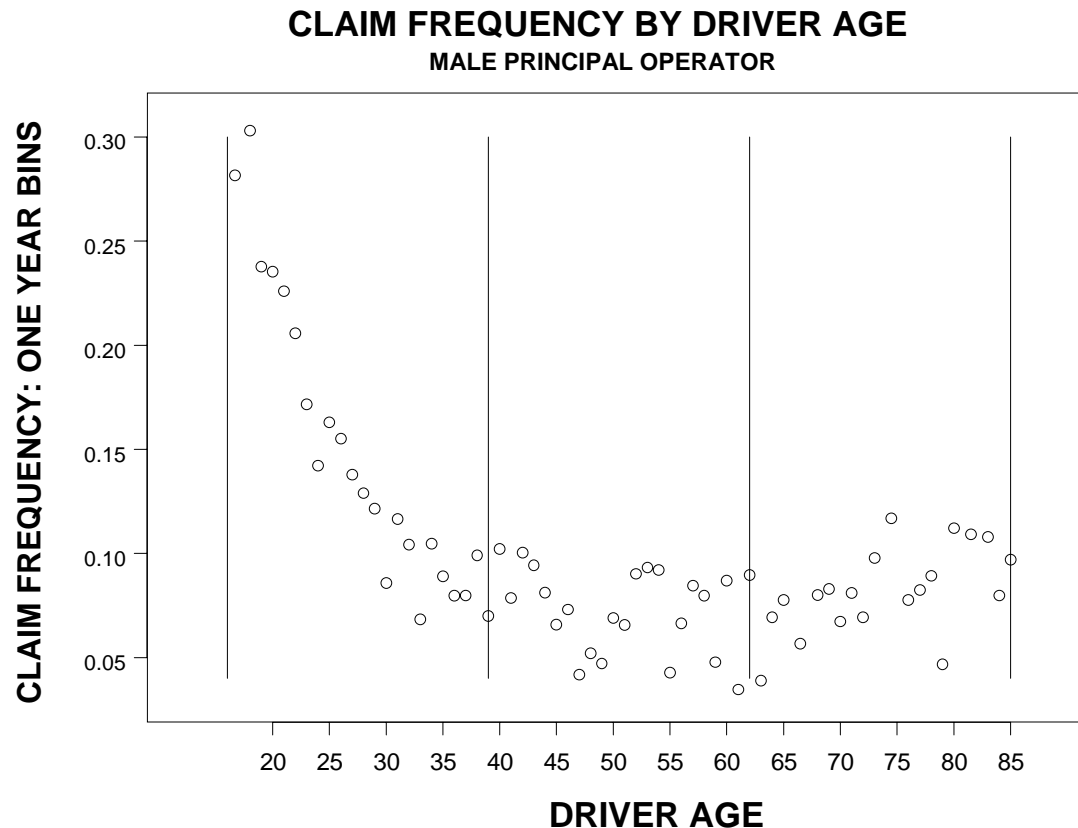
(Continued)

Nonlinearity – Polynomial

- Primary advantages
 - Can generally produce a good fit to a curved pattern
 - Model has fewer parameters than binning
- Primary disadvantages
 - More conceptually complicated than binning
 - Extrapolation can produce unrealistic projections
 - Difficult to modify shape of curve

Nonlinearity – Spline

Design Matrix
Nonlinearity
Missing Data



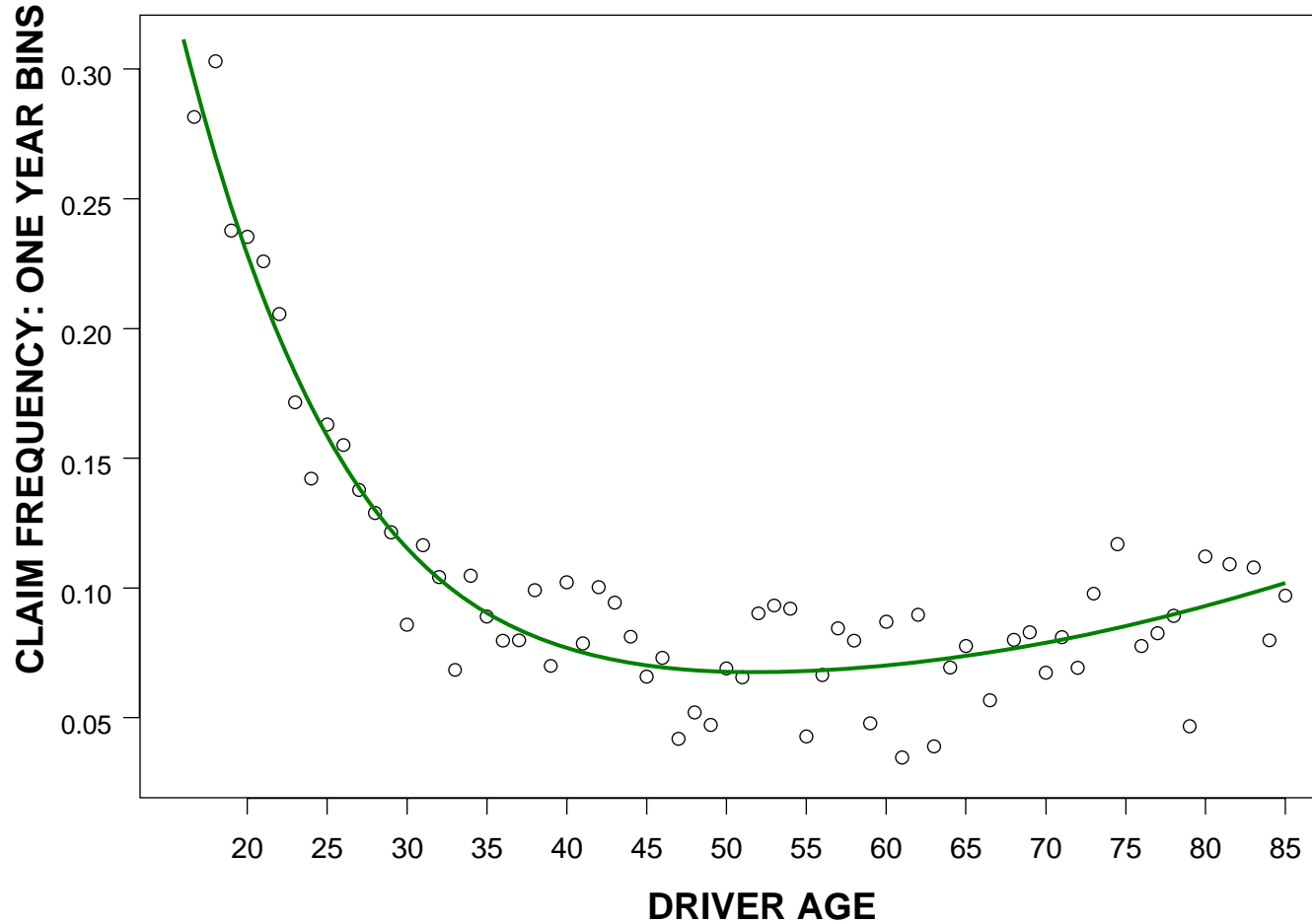
- Third degree polynomial between the knots
- Continuous value, first and second derivative at the knots
- Linear outside of the boundary knots

(Continued)

Nonlinearity – Spline

Design Matrix
Nonlinearity
Missing Data

PREDICTED CLAIM FREQUENCY - SPLINE
MALE PRINCIPAL OPERATOR



(Continued)

Design Matrix with Spline

Design Matrix
Nonlinearity
Missing Data

$$\begin{array}{cccc} \text{Intercept} & \text{Basis-1} & \text{Basis-2} & \text{Basis-3} \\ \left[\begin{array}{c} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{array} \right] & \left[\begin{array}{c} -.03 \\ .00 \\ -.05 \\ -.01 \\ -.11 \\ .47 \end{array} \right] & \left[\begin{array}{c} .44 \\ .00 \\ .28 \\ .05 \\ .46 \\ .36 \end{array} \right] & \left[\begin{array}{c} .59 \\ .00 \\ -.21 \\ -.03 \\ .65 \\ -.10 \end{array} \right] \end{array} \mathbf{X} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \end{bmatrix} = \begin{bmatrix} LP_1 \\ LP_2 \\ LP_3 \\ LP_4 \\ LP_5 \\ LP_6 \end{bmatrix}$$

- Three columns are added to the design matrix
 - These columns represent the spline basis
- GLM is fit with same likelihood and link functions

(Continued)

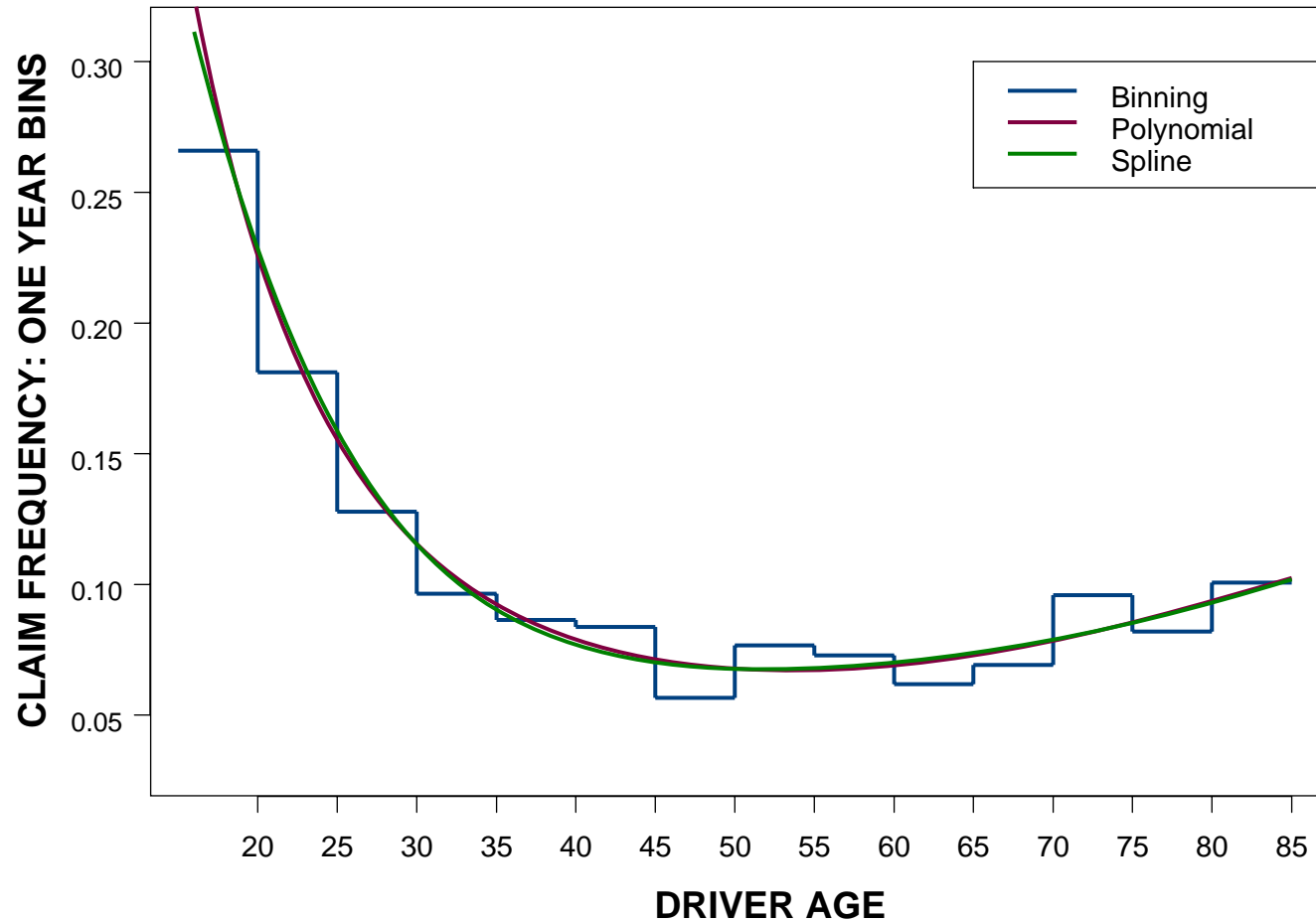
Nonlinearity – Spline

- Primary advantages
 - Can generally produce a good fit to a curved pattern
 - Model has fewer parameters than binning
 - More reasonable extrapolation than polynomial
 - Ability to manipulate shape of spline
- Primary disadvantage
 - More conceptually complicated than orthogonal polynomial

Comparison of Methods

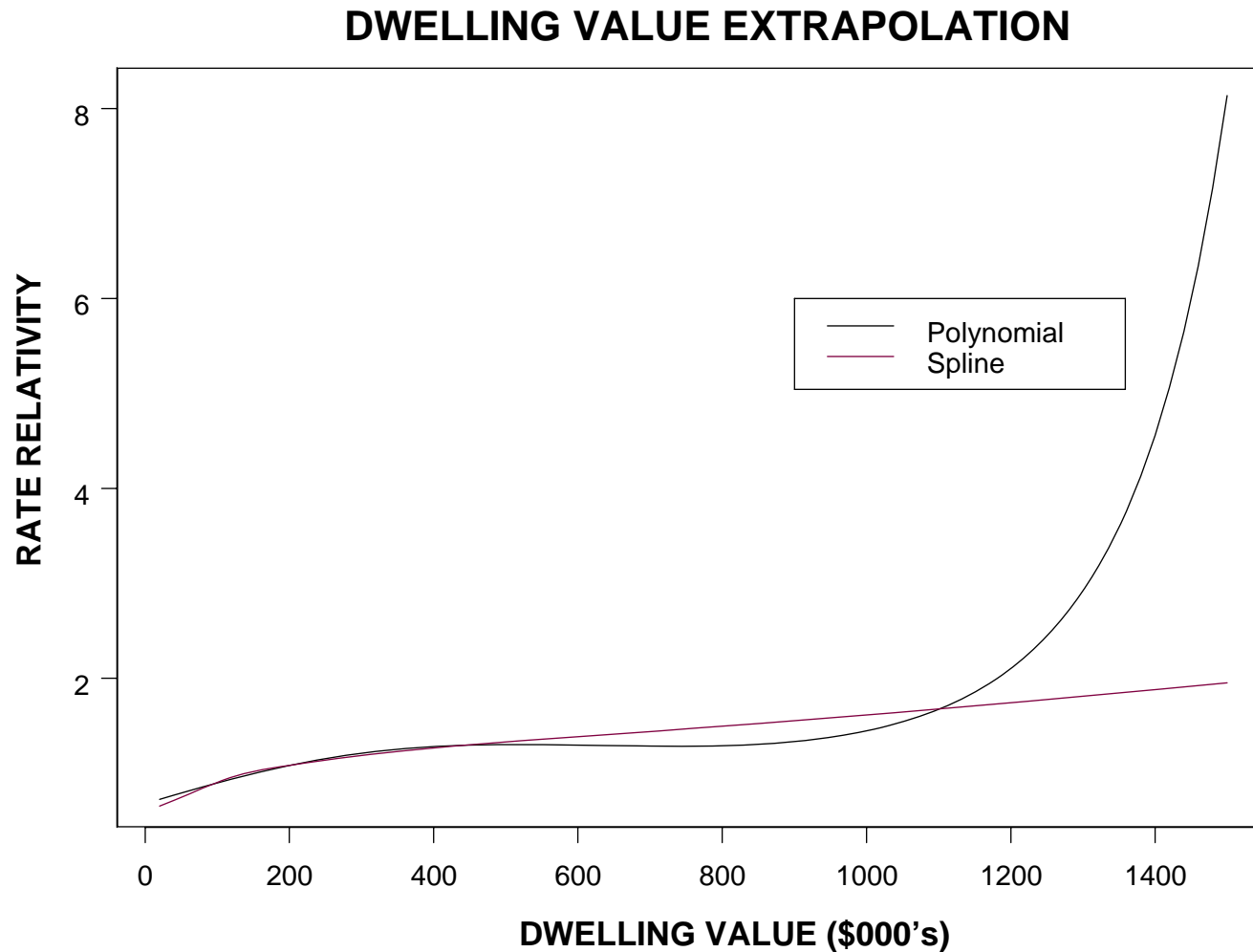
Design Matrix
Nonlinearity
Missing Data

PREDICTED CLAIM FREQUENCY
MALE PRINCIPAL OPERATOR



Extrapolation Example

Design Matrix
Nonlinearity
Missing Data



Missing Data – Description of Issue

Design Matrix
Nonlinearity
Missing Data

- Missing data can present unique challenges in model creation

Data

Design Matrix

<u>Class</u>	<u>State</u>	<u>AOI</u>	<u>Pop Density</u>	<u>Intercept</u>	<u>Class</u>	<u>ST MA</u>	<u>AOI</u>	<u>Pop Density</u>
65198	MA	125	.033	1	0	0	125	.033
65198	IL	235	.032	1	0	0	235	.032
70446	MA	240	.034	1	1	0	240	.034
70446	FL	350	.044	1	1	0	350	.044
64446	MA	100	.023	1	0	1	100	.023
64446	IN	110	NA	1	0	1	110	NA

(Continued)

Missing Data – Description of Issue

Design Matrix
Nonlinearity
Missing Data

- What methodologies exist for addressing missing data?

Intercept	Class		ST MA	AOI	Pop Density
1	0	0	1	125	.033
1	0	0	0	235	.032
1	1	0	1	240	.034
1	1	0	0	350	.044
1	0	1	1	100	.023
1	0	1	0	110	NA

(Continued)

Missing Data – Methodology #1

Design Matrix
Nonlinearity
Missing Data

- Listwise deletion: Eliminate any row in the design matrix with missing values

Intercept	Class		ST MA	AOI	Pop Density
1	0	0	1	125	.033
1	0	0	0	235	.032
1	1	0	1	240	.034
1	1	0	0	350	.044
1	0	1	1	100	.023

(Continued)

Missing Data – Methodology #2

Design Matrix
Nonlinearity
Missing Data

- Mean imputation: Replace missing values with mean of values where data is present

Intercept	Class		ST MA	AOI	Pop Density
1	0	0	1	125	.033
1	0	0	0	235	.032
1	1	0	1	240	.034
1	1	0	0	350	.044
1	0	1	1	100	.023
1	0	1	0	110	.033

(Continued)

Missing Data – Methodology #3

Design Matrix
Nonlinearity
Missing Data

- Linear mean imputation: Create spline basis excluding missing values and mean impute on spline basis

Intercept	Class		ST MA	AOI	Pop Density		
1	0	0	1	125	.033	.109	.359
1	0	0	0	235	.032	.102	.328
1	1	0	1	240	.034	.116	.393
1	1	0	0	350	.044	.194	.852
1	0	1	1	100	.023	.053	.122
1	0	1	0	110			

(Continued)

Missing Data – Methodology #3

Design Matrix
Nonlinearity
Missing Data

- Linear mean imputation: Create spline basis excluding missing values and mean impute on spline basis

Intercept	Class		ST MA	AOI	Pop Density		
1	0	0	1	125	.033	.109	.359
1	0	0	0	235	.032	.102	.328
1	1	0	1	240	.034	.116	.393
1	1	0	0	350	.044	.194	.852
1	0	1	1	100	.023	.053	.122
1	0	1	0	110	.033	.411	.115

(Continued)

Missing Data – Methodology #4

Design Matrix
Nonlinearity
Missing Data

- Single imputation: Use other predictor variables to build a model and impute missing values
 - Example: Model Pop Density based on AOI

Intercept	Class		ST MA	AOI	Pop Density
1	0	0	1	125	.033
1	0	0	0	235	.025
1	1	0	1	240	.034
1	1	0	0	350	.044
1	0	1	1	100	.023
1	0	1	0	110	.027

(Continued)

Missing Data – Methodology #5

Design Matrix
Nonlinearity
Missing Data

- Multiple imputation: Use other predictor variables to model missing values
 - Multiple imputations are created based on distribution of residuals in estimates of missing values

Intercept	Class		ST MA	AOI	Pop Density
1	0	0	1	125	.033
1	0	0	0	235	.025
1	1	0	1	240	.034
1	1	0	0	350	.044
1	0	1	1	100	.023
1	0	1	0	110	.027

AOI	Pop Density
125	.033
235	.025
240	.034
350	.044
100	.023
110	.029

AOI	Pop Density
125	.033
235	.025
240	.034
350	.044
100	.023
110	.025

ST MA	Pop Density
1	.033
0	.025
1	.034
0	.044
1	.023
0	.025

Steps in Multiple Imputation Process

1. Choose starting values for mean and covariance matrix of predictor variables
2. Use mean and covariance matrix to estimate regression parameters
3. Use regression parameters to estimate missing values. Add a random draw from the residual normal distribution for that variable
4. Use the resulting data set to compute new mean and covariance matrix
5. Make a random draw from the posterior distribution of the means and covariances
6. Use the random draw from step five, go back to step two and cycle through the process until convergence is achieved

(Continued)

Steps in Multiple Imputation Process

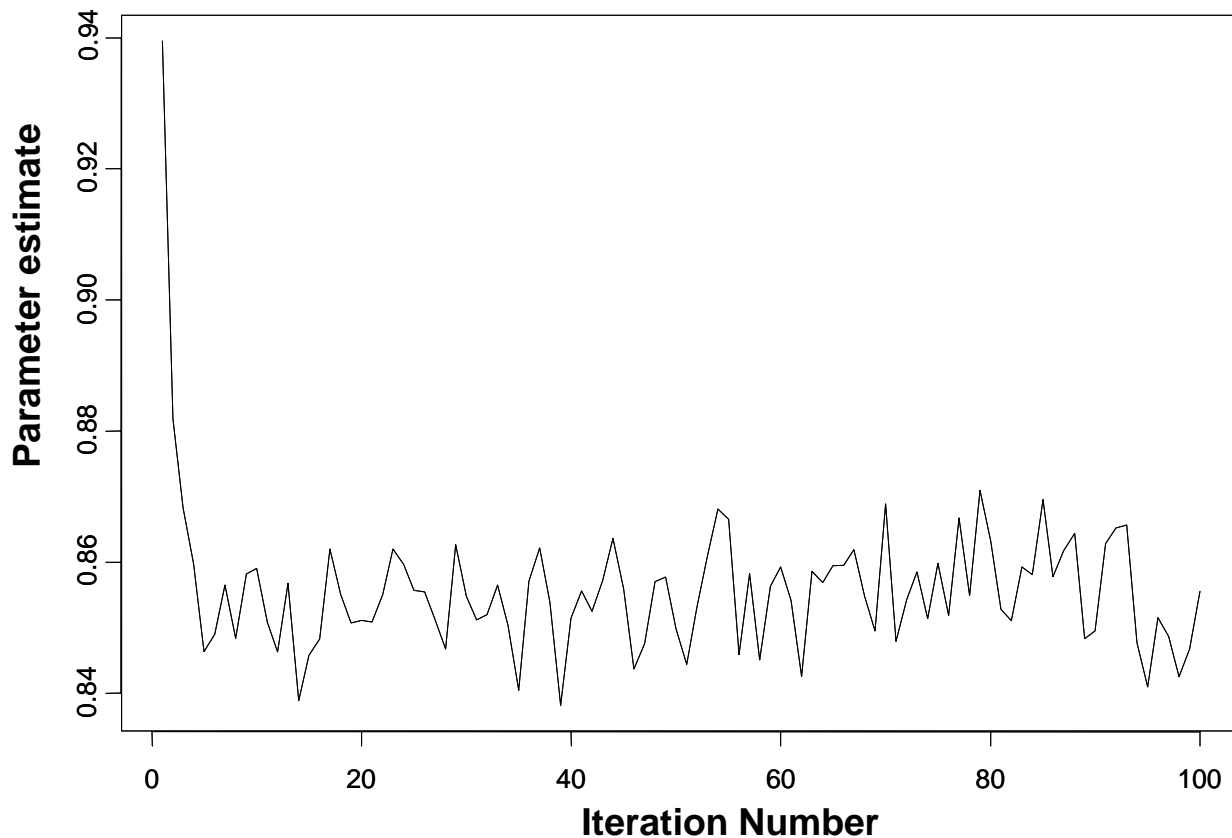
- Assumptions underlying multiple imputation algorithms
 - Data is missing at random: Missingness of predictor variable “V” cannot depend on value of “V” but can depend on values of other predictor variables
 - Data is distributed with a multivariate normal distribution
- Two issues that must be addressed
 - Initial convergence of iterations
 - Correlation of consecutive iterations

(Continued)

Time Series Plot

Design Matrix
Nonlinearity
Missing Data

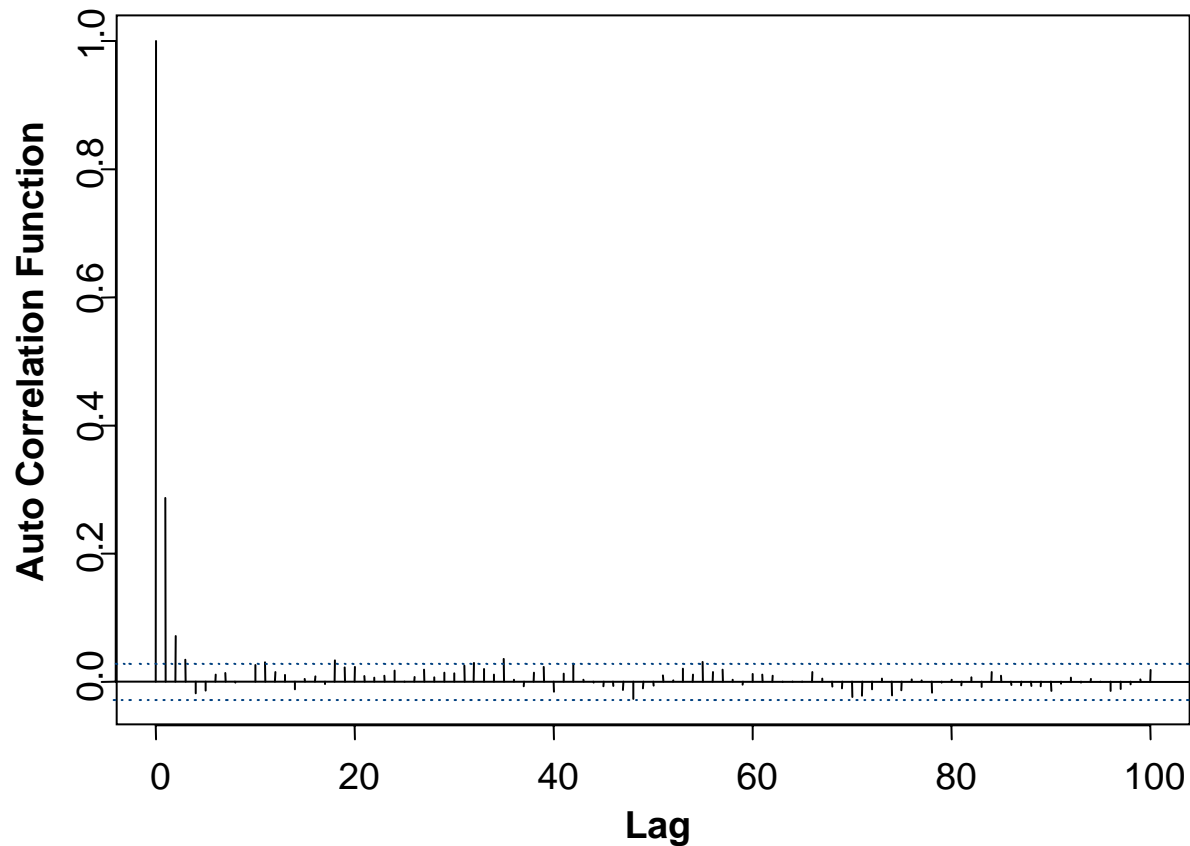
- Initial convergence is assessed via a time series plot



(Continued)

Autocorrelation Plot

- Spread between iterations is assessed via an autocorrelation plot



Testing of Missing Value Methods

Design Matrix
Nonlinearity
Missing Data

- Method #1
 - Created both training and holdout data sets
 - Both contained missing data
 - Built models of claim frequency under different missing value analysis methods with training data set
 - Identical predictor variables in all models
 - Compared results (deviance) of methods in holdout data set where all data is present

(Continued)

Testing of Missing Value Methods

- Method #2
 - Created a model of missing probability
 - Limited modeling database to observations in which all data was present
 - Randomly generated missing values based on missing probability
 - 100 iterations
 - Built models of claim frequency under different missing value analysis methods
 - Identical predictor variables in all models
 - Compared results (deviance) of methods in data set where all data is present

Ranking the Performance of Missing Value Methods

1. Single imputation/Multiple imputation
2. Linear mean imputation
3. Mean imputation
4. Listwise deletion

Missing Data Framework

Design Matrix
Nonlinearity
Missing Data

- Questions

- What is the level of missing data?
- What can be inferred about the missing data mechanism?
- What is the size of the modeling database in which all values are present?
- Will the data continue to be missing when the model is applied?

(Continued)

Missing Data Framework

Design Matrix
Nonlinearity
Missing Data

- Actions

- For low proportions of missing data: Listwise deletion
- For higher proportions of missing data in a large modeling database: Listwise deletion with oversampling
- For mid-to-small modeling databases: Employ imputation
 - Initial exploration with linear mean imputation
 - Fit final model with single imputation or multiple imputation

Sources

- Orthogonal Polynomials
 - Wolfram Mathworld: <http://mathworld.wolfram.com/Gram-SchmidtOrthonormalization.html>
- Splines
 - Hastie, Tibshirani and Friedman: *The Elements of Statistical Learning*
- Missing Data
 - Paul Allison: *Missing Data*
 - J.L. Schafer: *Analysis of Incomplete Multivariate Data*
 - Insightful Corporation: *Analyzing Data with Missing Values in S-Plus*