More Flexible GLMs
2009 CAS Ratemaking and Product Management Seminar

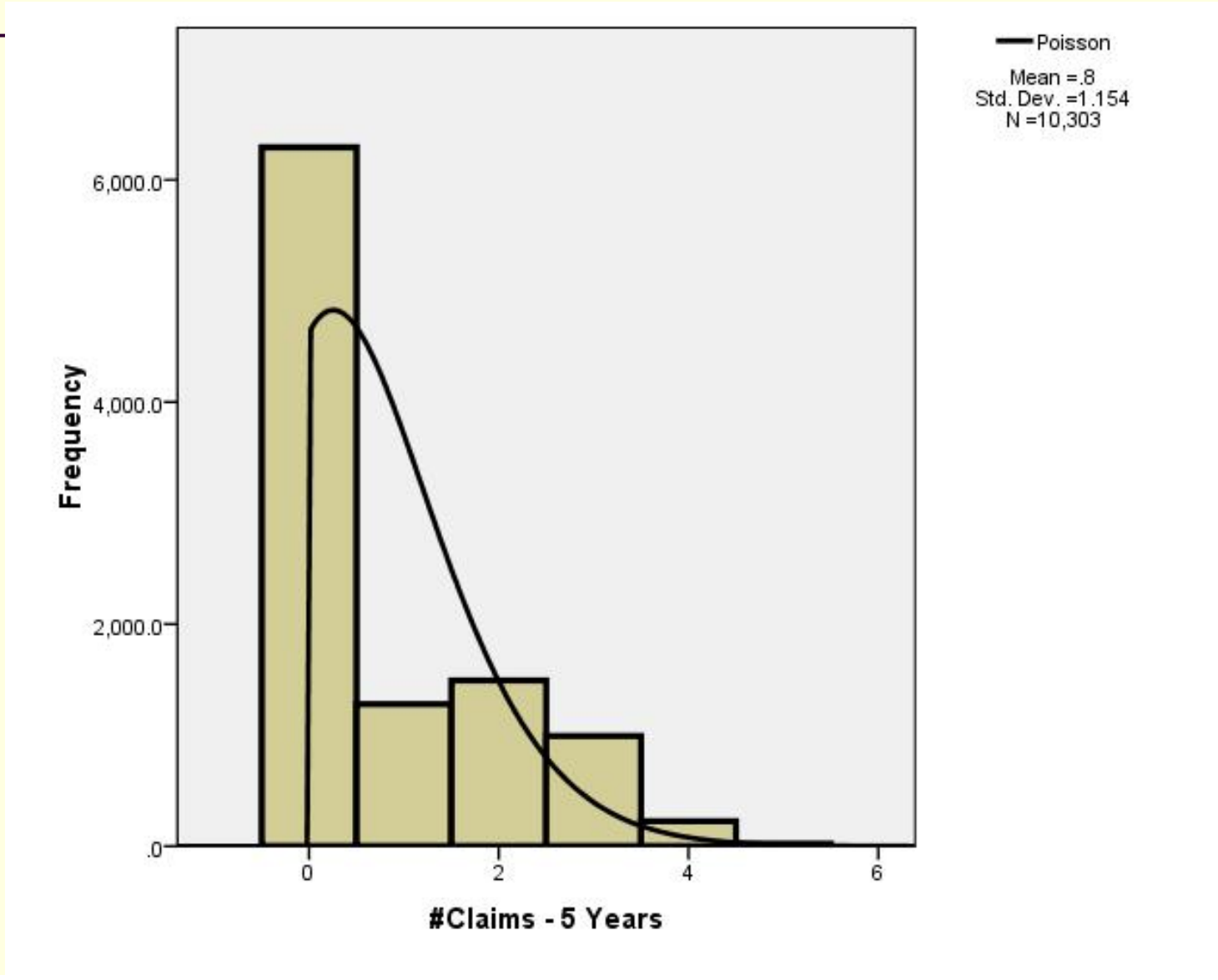
Matt Flynn, PhD
Louise Francis, FCAS, MAAA
Louise_francis@msn.com
www.data-mines.com

Address Two Modeling issues

- Adjusted model for accidents

- Use of tree Models

Zero-Adjusted Distributions



Poisson

$$f(y,\lambda)=\frac{e^{-\lambda}\lambda^{y}}{y!}\quad y = 0, 1, 3, \ldots$$

mean = λ

var = λ

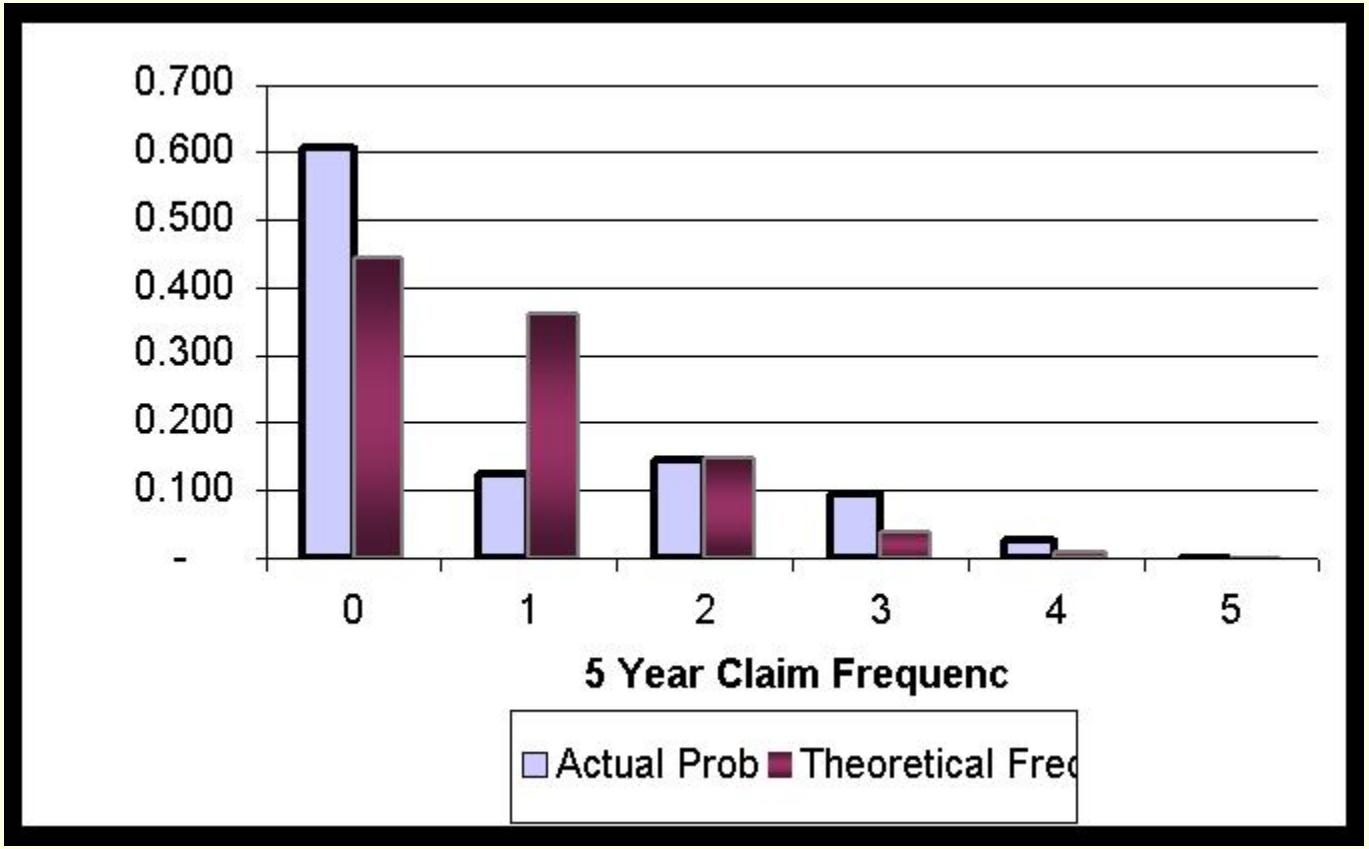
ll=y*log$(\lambda)-\lambda-\lg amma(y+1)$

http://en.wikipedia.org/wiki/Poisson_distribution

http://mathworld.wolfram.com/PoissonDistribution.html

http://www.itl.nist.gov/div898/handbook/eda/section3/eda366j.htm

Actual vs. Poisson Frequencies



Negative Binomial

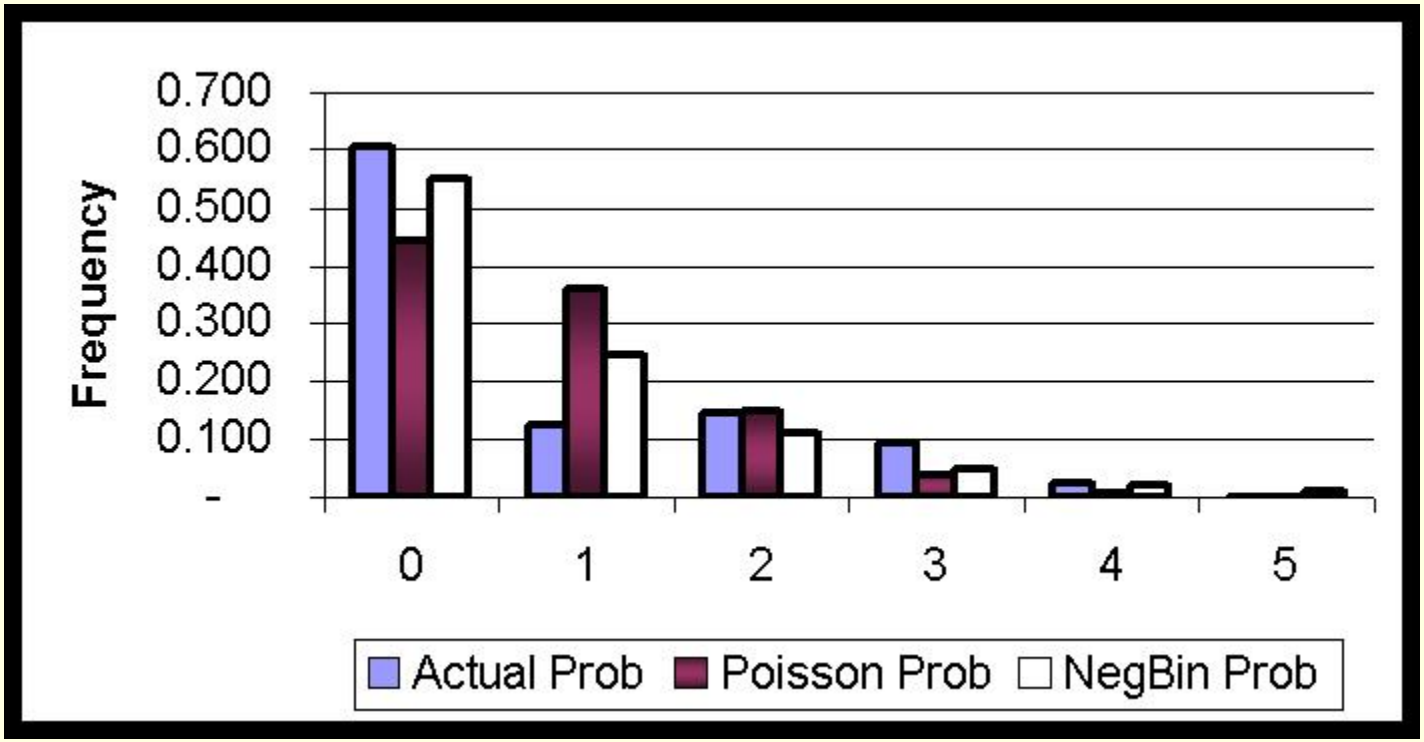
$$f(k) = \frac{\Gamma(r+k)}{k!\Gamma(r)} p^r (1-p)^k$$

$$mean = r\frac{1-p}{p}$$

$$\text{var}\,iance = r\frac{1-p}{p^2}$$

http://en.wikipedia.org/wiki/Negative_binomial_distribution

Comparison of Actual, Poisson and Negative Binomial Frequencies



Geometric

$$f(k) = \frac{\Gamma(r+k)}{k!\Gamma(r)} p^r (1-p)^k$$

$$mean = r\frac{1-p}{p}$$

$$\text{var}\,iance = r\frac{1-p}{p^2}$$

http://en.wikipedia.org/wiki/Negative_binomial_distribution

Zero-Inflated Poisson (ZIP)

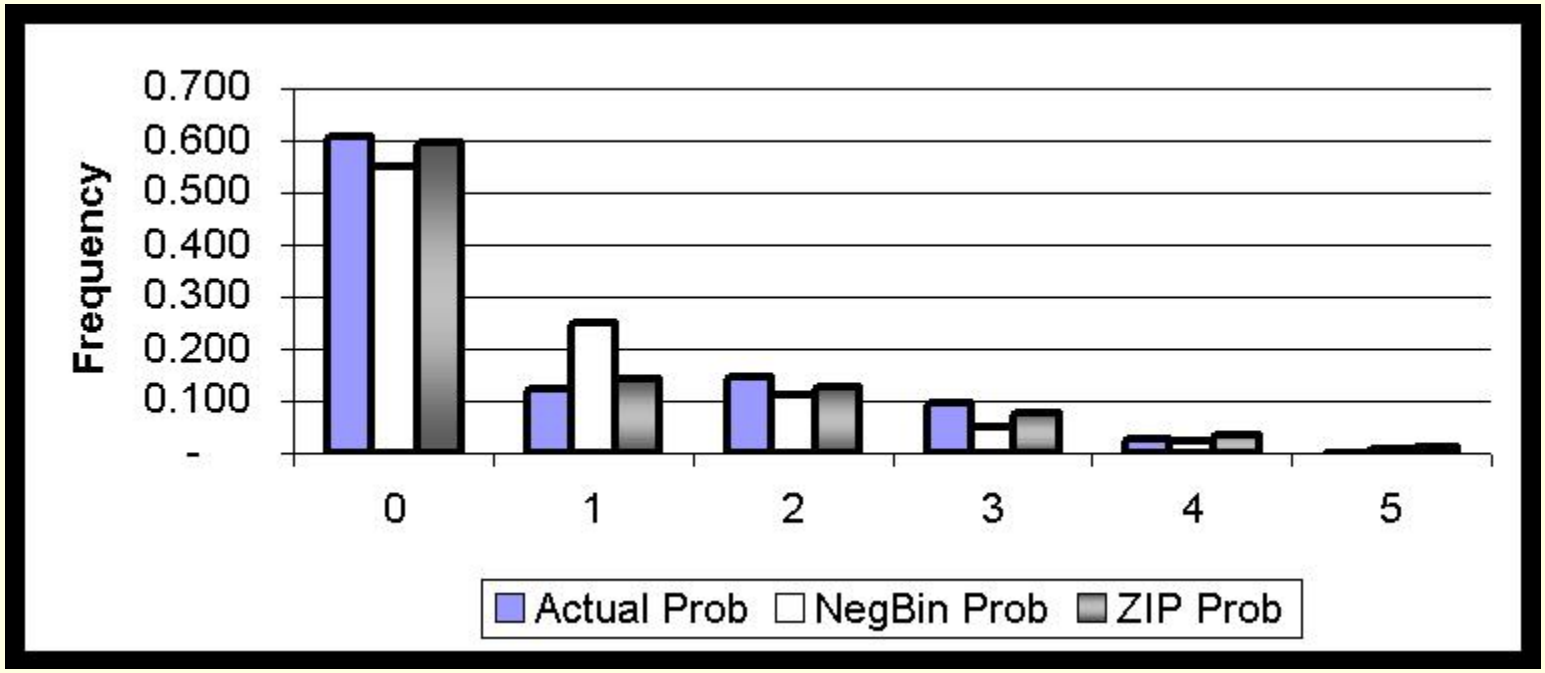
$$\varphi + (1-\varphi)e^{-\lambda} \quad x = 0$$

$$(1-\varphi)\frac{\lambda^x}{x!}e^{-\lambda} \quad x > 0$$

$$mean = \mu(1-\varphi)$$

$$\text{var}\,iance = \mu(1-\varphi)(1+\mu\varphi)$$

Actual, Negative Binomial and Zero Inflated Poisson Frequencies



Zero-Inflated Negative Binomial (ZINB)

$$\varphi + (1-\varphi)NB(0,r,p)_{x=0}$$

$$(1-\varphi)NB(x,r,p)_{x>0}$$

$$mean = \mu(1-\varphi)$$

$$\text{var}\,iance = \mu(1-\varphi)(1+\mu(\varphi+\alpha))$$

Test for Excess zeros

$$S = \frac{\left\{ \sum_{i=1}^{n} (I(x_i = 0) - p_{0i}) / p_{0i} \right\}^2}{\sum_{i=1}^{n} (I(x_i = 0) - p_{0i}) / p_{0i} - n\bar{x}}$$

Gini Index

$$A = \left(\frac{1}{2} F_1 + F_{2\ldots} + F_{n-1} + \frac{1}{2} F_n \right) * \Delta x,$$

F_i is ith cumulative income

Gini Index Result

Treatment of Variable	Poisson	ZIP
Original Variables	0.1770	0.1830
CHAID, MV Capped	0.1780	0.1800
CHAID, MV Binned	0.1760	0.1800

GLM Regression Model

- Y = η+e = **x'β+e**
- a random component, denoted e
- a linear relationship between a dependent variable and its predictors

$$\eta = a + b_1 x_1 + b_2 x_{2+} \ldots b_n x_n$$

- a link function

The Claim Model

- Y = f(car use, marital status, density, gender) + e
  - the identity link $\mu=\eta$
  - the log link $\mu=\exp(\eta)$ or $\eta=\log(\mu)$.

Comparison of Models

Model	-2*log-likelihood
Poisson	7,141.9
Overdispersed Poisson	6,843.9
Geometric	6,764.1
Negative Binomial	6,764.1
ZINB	6,541.2
ZIP	6,404.0

CHAID Trees

CART Example with Seven Nodes
IME Proportion as a Function of Provider Bill

Cross Tabulation of Count Data

Expected Count

Home/Work	Claim Indicator: No Claim	Claim Indicator: Claims	Total
Highly Rural	3,10.30	197.70	508
Highly Urban	2,198.20	1,400.80	3,599
Rural	955.90	609.10	1,565
Urban	2,828.60	1,802.40	4,631

Chi Squared Statistic: $(O-E)^2/E$

Home/Work	Claim Indicator: No Claim	Claim Indicator: Claims
Highly Rural	64.70	101.60
Highly Urban	98.90	155.20
Rural	178.50	280.20
Urban	2.80	4.40

886.20

Chi Square Statistic

$$\chi_{k-1}^2 = \frac{(Observed - Fitted)^2}{Fitted}$$

Frequencies

Percent of Policies With Claims			
		Claim Indicator	
		No Claim	Claims
Home/Work	Highly Rural	89%	11%
	Rural	87%	13%
	Urban	59%	41%
	Highly Urban	48%	52%
	Total	61%	39%

Consolidate Categories

Observed			
	No Claim	Claim	Total
Highly Rural	452	56	508
Rural	1,369	196	1,565
Total	1,821	252	2,073

Expected			
	No Claim	Claims	Total
Highly Rural	446.25	61.75	508
Rural	1,374.75	190.25	1,565

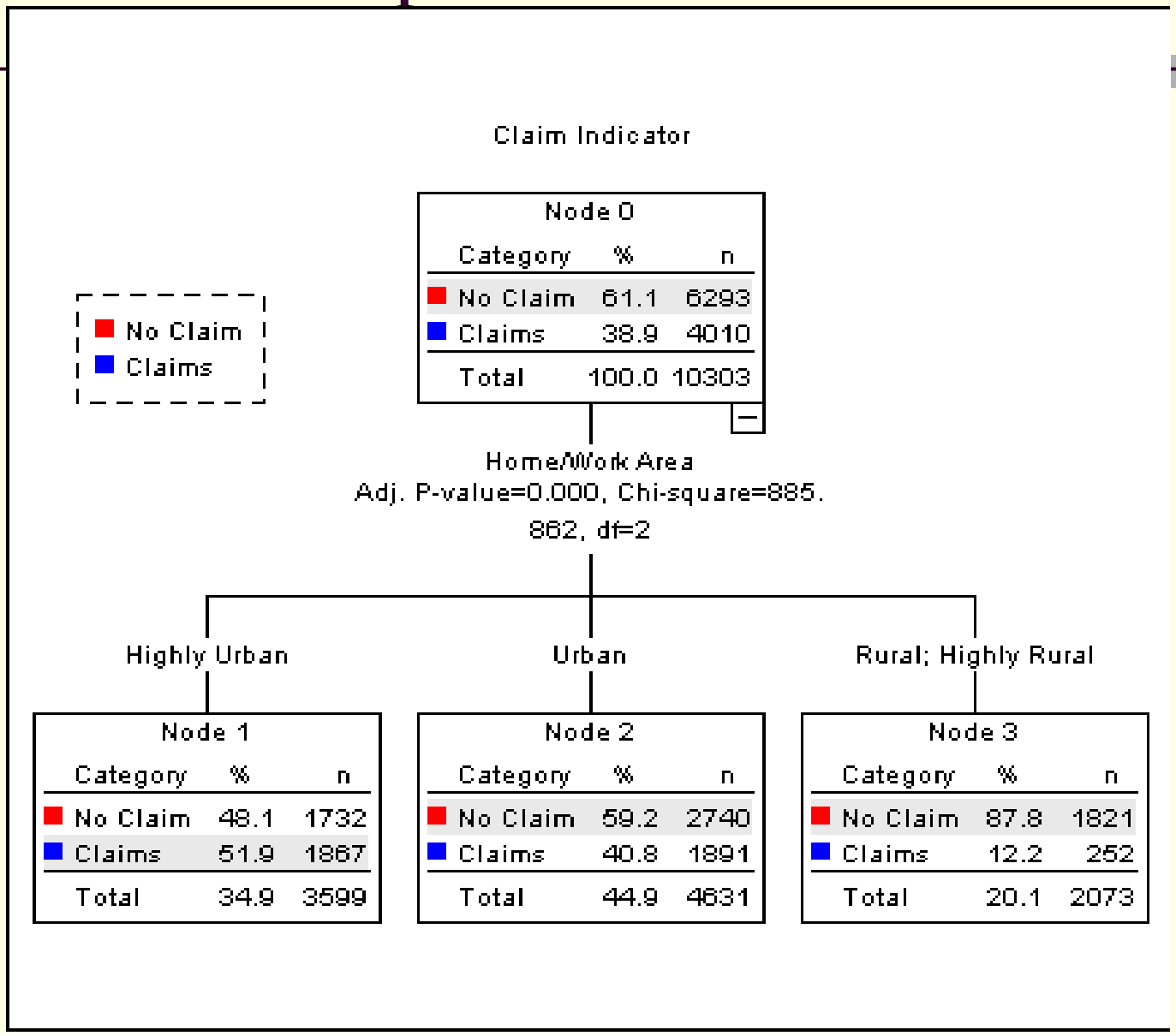
Chi Squared		
	No Claim	Claims
Highly Rural	0.07	0.54
Rural	0.02	0.17
Total	0.09	0.81

Consolidate Results

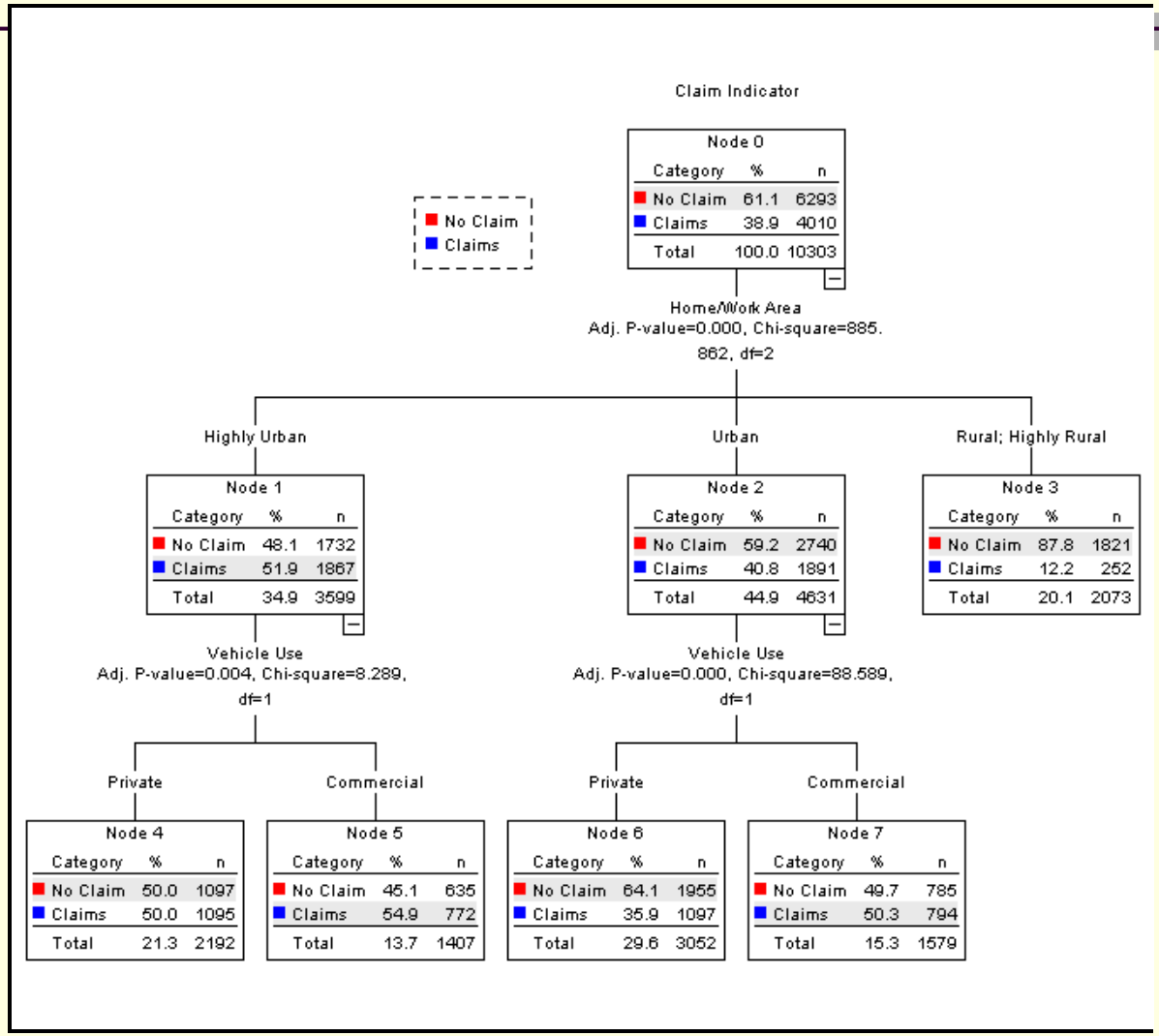
Area * Claim Indicator Crosstabulation			
	Claim Indicator		
	No Claim	Claims	Total
Rural	1,821	252	2,073
Urban	2,740	1,891	4,631
Highly Urban	1,732	1,867	3,599
Total	6,293	4,010	10,303

Percent of Policies With A Claim			
	Claim Indicator		
	No Claim	Claims	Total
Rural	88%	12%	508
Urban	59%	41%	3,599
Highly Urban	48%	52%	1,565

Tree Graphic



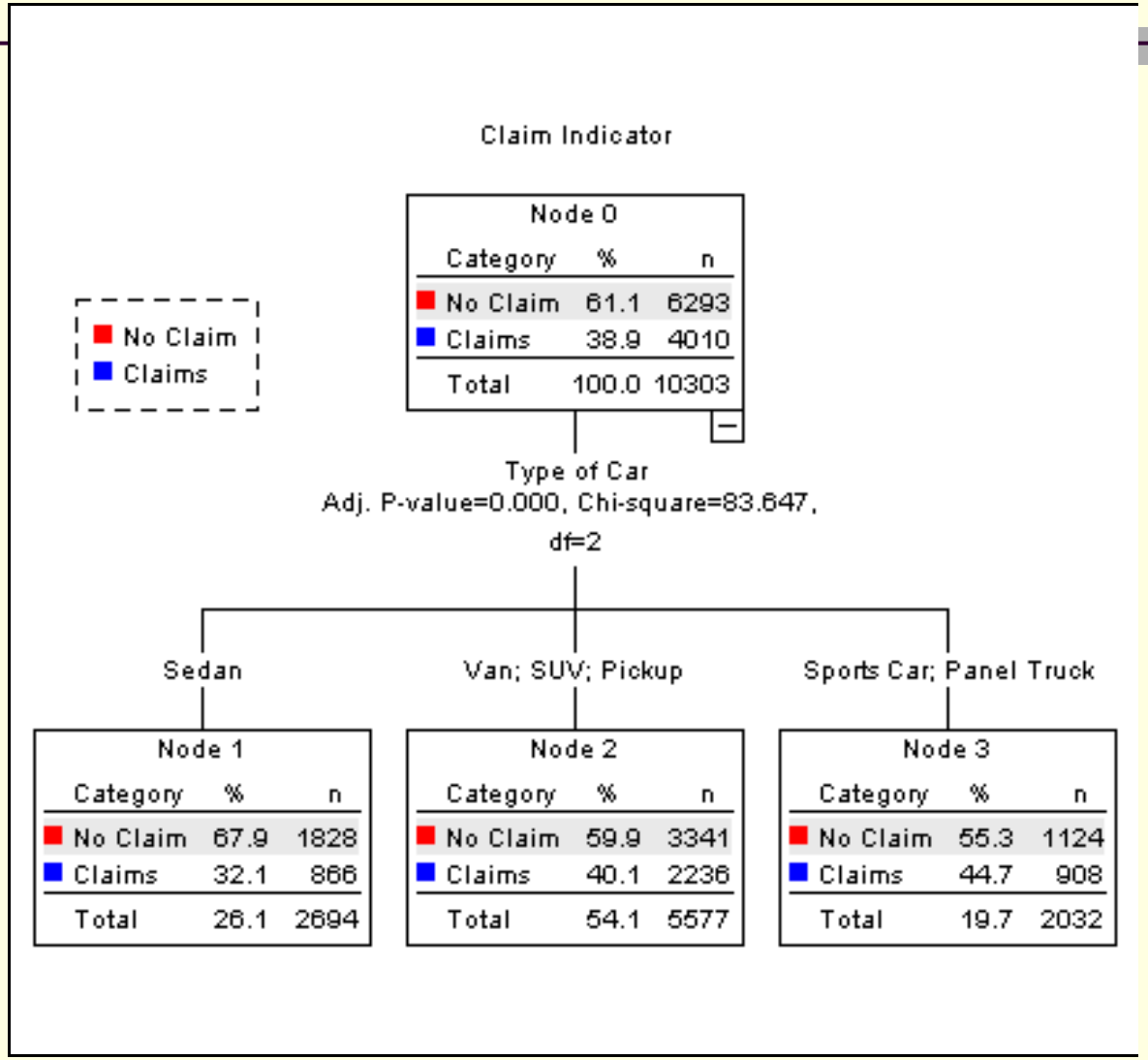
Tree With Two Variables



Significant Differences by Type?

Car Type	Frequency	Percent
Panel Truck	853	8%
Pickup	1,772	17%
Sedan	2,694	26%
Sports Car	1,179	11%
SUV	2,883	28%
Van	922	9%
Total	10,303	100%

CHAID Tree for Car Type



Continuous Variable

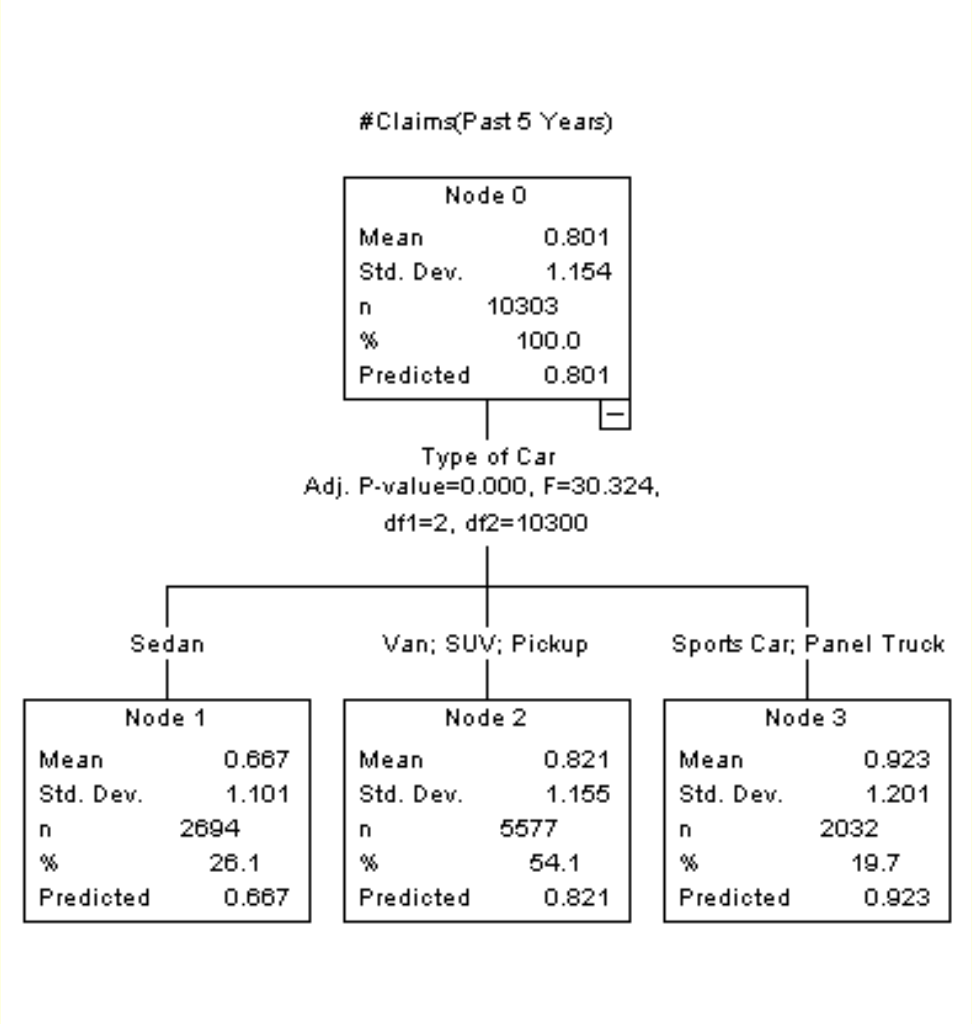
$$F = \frac{(RSS_1 - RSS_2)/(p_2 - p_1)}{RSS_2 / p_2}$$

RSS = residual sum of squares

p_1 = degrees of freedom for model 1

p_2 is degrees of freedom for model 2

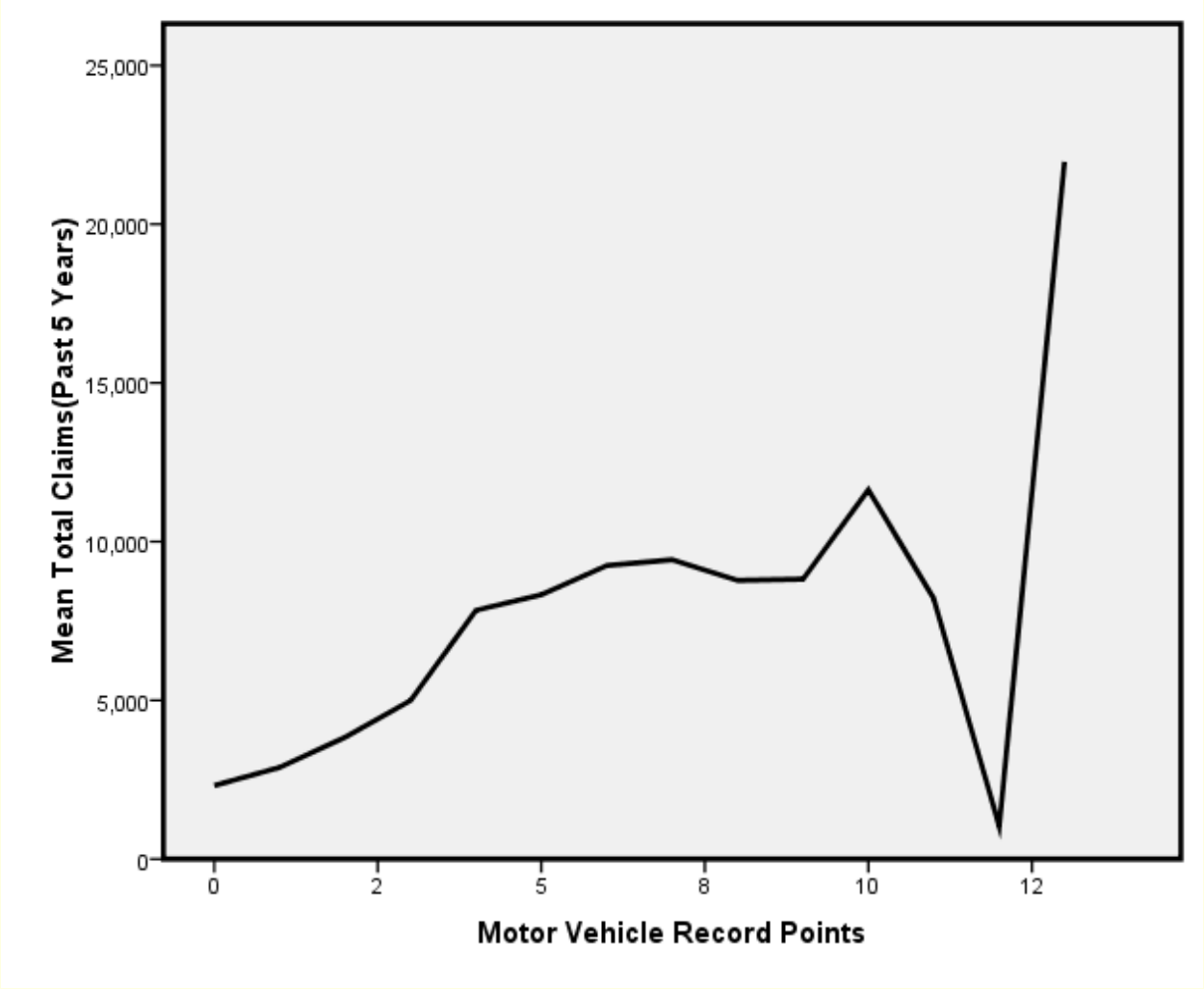
Regression Tree



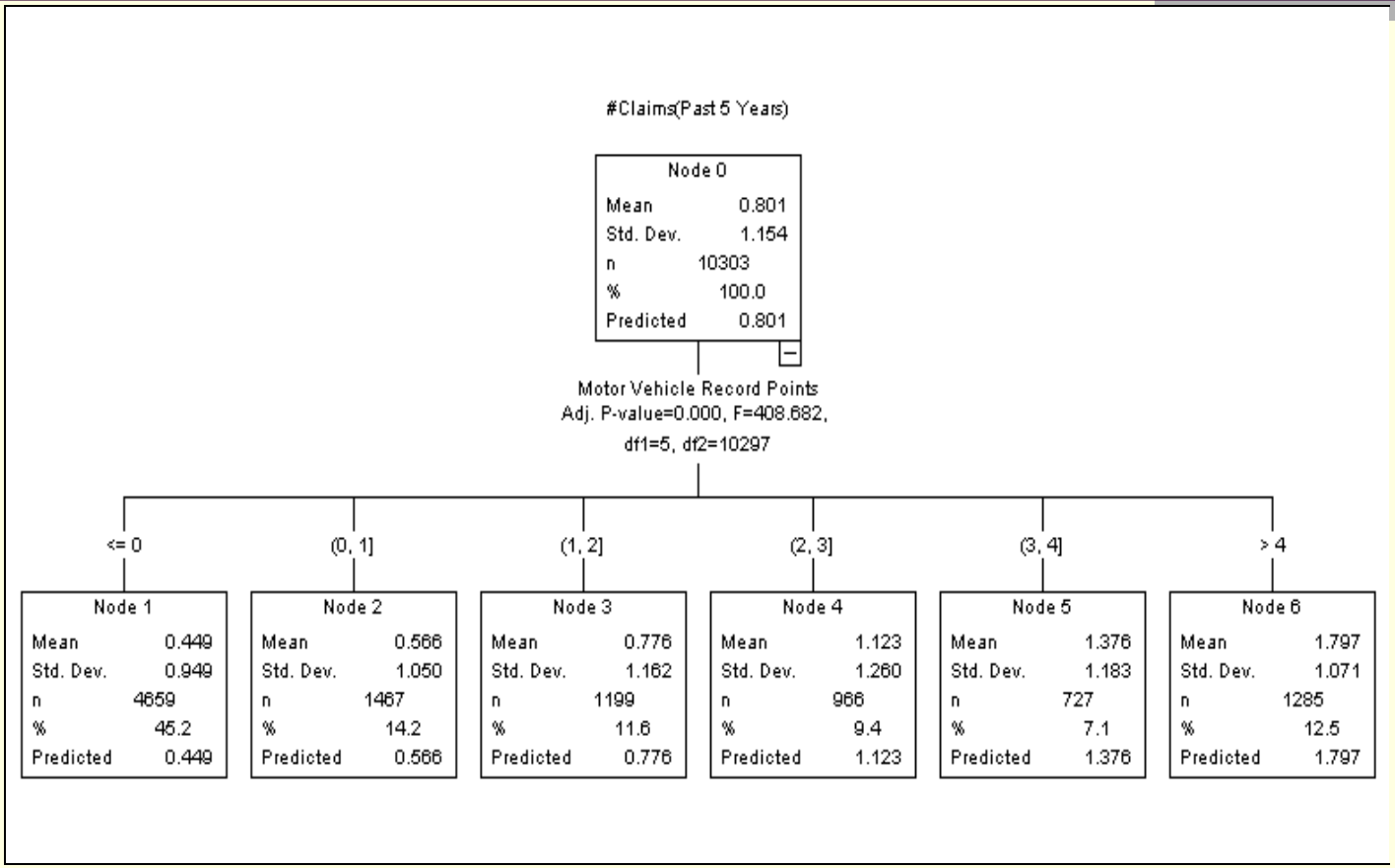
Akaike Information

	Original Variables	Reduced Variables
Poisson Regression	12,066	12,026
ZIP	12,006	12,020

Motor Vehicle Points



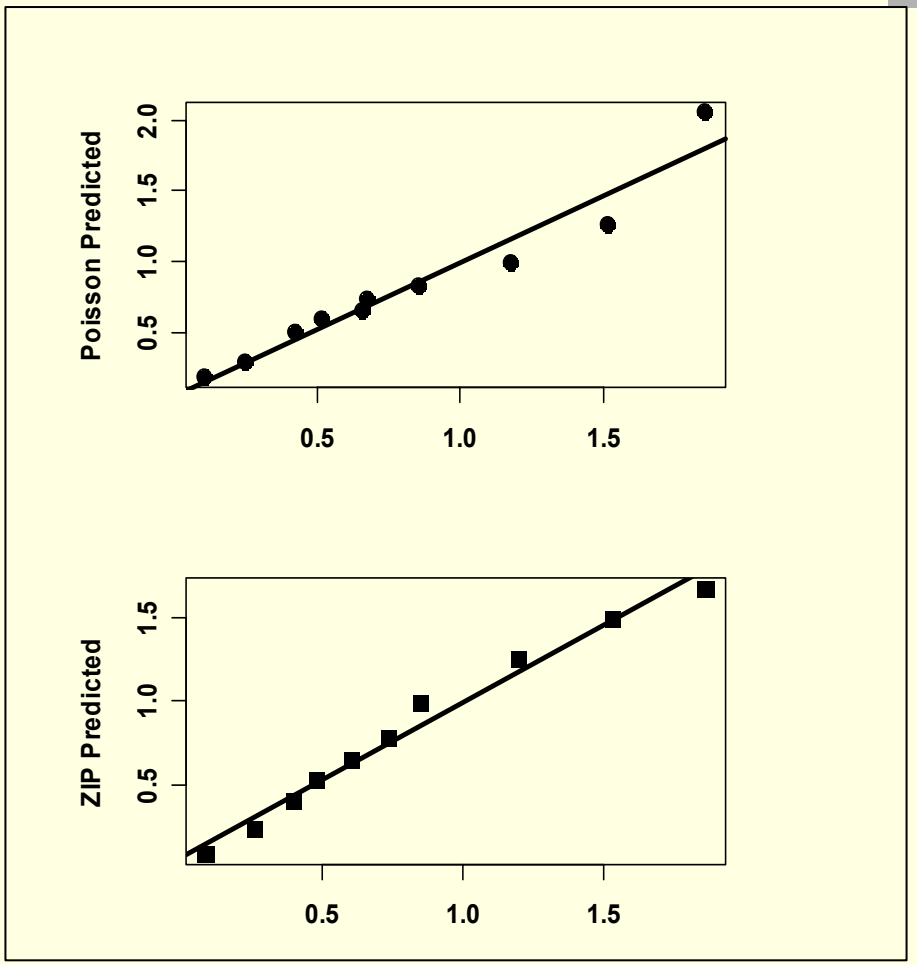
Tree for MV Points



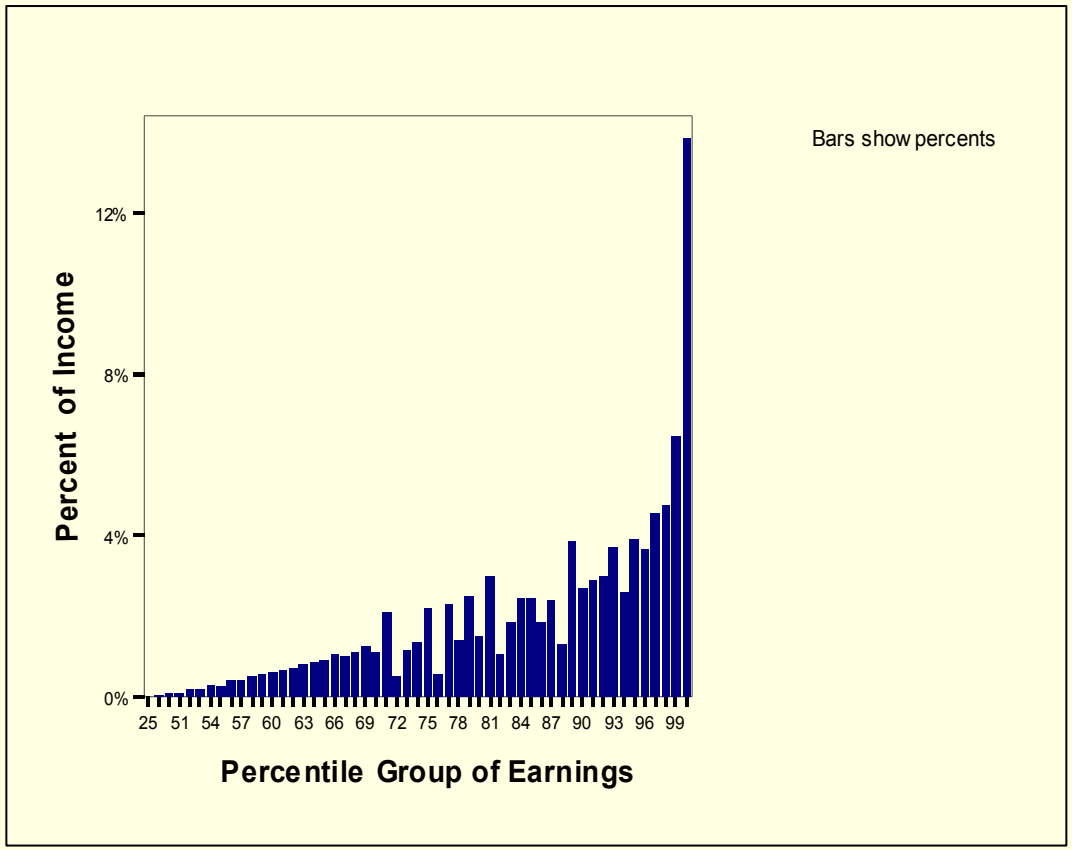
AIC

Treatment of Variable	Poisson	ZIP
MV Points	12,593	11,022
Capped MV Points	12,502	11,066
Binned MV Points	12,496	10,946

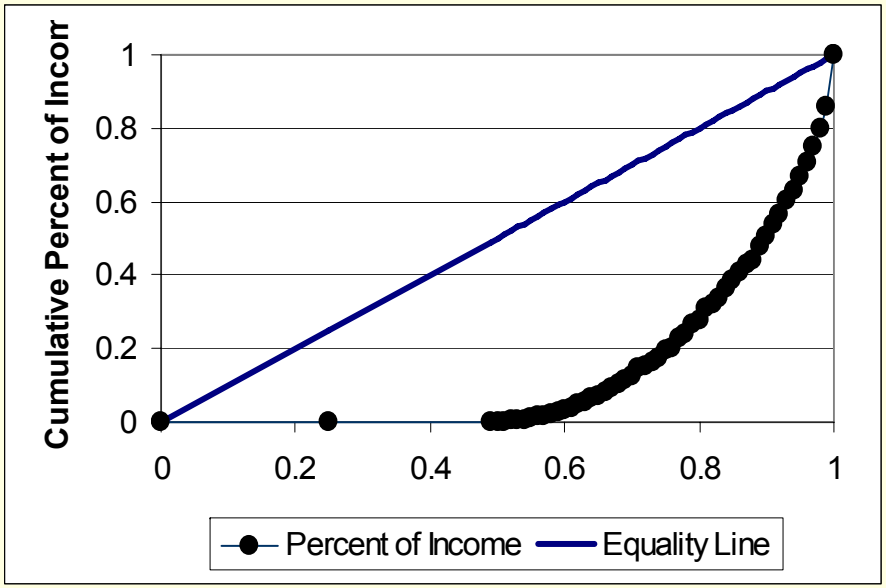
Goodness of Fit



Lorenz Curve



Lorenz Curve



Gini Results

Treatment of Variable	Poisson	ZIP
Original Variables	0.1770	0.1830
CHAID, MV Capped	0.1780	0.1800
CHAID, MV Binned	0.1760	0.1800

Conclusions

- Zero adjusted distribution may fit claim frequency data better than Poisson or negative binomial
- CHAID and other partitioning techniques can be helpful in preprocessing data for category reduction