More Flexible GLMs
Zero-Inflated Models and Hybrid Models

Mathew Flynn, Ph.D.

Louise A. Francis FCAS, MAAA

Motivation: GLMs are widely used in insurance modeling applications. Claim or frequency models are a key component of many GLM ratemaking models. Enhancements to the traditional GLM that are described in this paper may by able to address practical issues that arise when fitting count models to insurance claims data.

For modeling claims within the GLM framework, the Poisson distribution is a popular distribution choice. In the presence of overdispersion, the negative binomial is also sometimes used. The statistical literature has suggested that taking excess zeros into account can improve the fit of count models when overdispersion is present. In insurance excess zeros may arise when claims near the deductible are not reported to the insurer, thus inflating the number of zero policies when compared to the predictions of a Poisson or Negative Binomial distribution.

In predictive modeling practice, data mining techniques such as neural networks and decision trees are often used to handle data complexities such as nonlinearities and interactions. Data mining techniques are sometimes combined with GLMs to improve the performance and/or efficiency of the predictive modeling analysis. One augmentation of GLMs uses decision tree methods in the data preprocessing step. An important preprocessing task reduces the number of levels on categorical variables so that sparse cells are eliminated and only significant groupings of the categories remain.

Method: This paper addresses some common problems in fitting count models to data. These are:

- Excess zeros
- Parsimonious reduction of category levels
- Nonlinearity

Results: The research described in this paper applied zero-inflated and hybrid models to claim frequency data. The research suggests that mixtures of GLM models incorporating adjustments for excess zeros improves the fit of the model compared to single distribution count models on some count data. The analysis also indicates that variable preprocessing using the CHAID tree technique can help reduce the complexity of models by retaining only category groupings that are significant with respect to their impact on the dependant variable.

Conclusions: By incorporating greater flexibility into GLM count models, practitioners may be able to improve the fit of models and increase the efficiency of the modeling effort. Use of the ZIP or ZINB improves the model fit for an illustrative automobile insurance database. The ZIP or ZINB distributions also provided a better overall approximation to the unconditional distribution of the data for the fit of a few additional insurance and non-insurance database. While the categorical variables in our illustrative data contained only a few categories compared to most realistic applications databases encountered in insurance, the fit of several predictive models. We also illustrate how the procedure can be applied to efficiently preprocess categorical variables with large numbers of categories.

Availability: Excel spreadsheets comparing the Poisson, negative binomial, zero-inflated Poisson and zero-inflated negative binomials well as R code for reproducing many models used in this paper will be available on the CAS Web Site.

Keywords: Predictive modeling, automobile ratemaking, generalized linear models, data mining

1. INTRODUCTION

Generalized linear models (GLMs) use a regression procedure to fit relationships between predictor and target variables. Unlike classical ordinary least squares regression where the random component (i.e., the error term) is assumed to follow a normal distribution, the random component in a GLM is assumed to belong to the exponential family of distributions. This family includes, along with the normal, the Poisson, the gamma and others commonly encountered in statistical analysis. GLMs are widely used in insurance modeling applications. In both the classical statistical literature (McCullagh and Nelder, 1989) and insurance-specific literature (de Jong and Heller, 2008) GLM techniques are applied to modeling insurance frequency and severity data. GLMs are a linear modeling procedure, since the relationship between a suitable transform of the dependent variable and the independent variables is assumed to be linear.

Commonly used data mining techniques employ automated procedures to efficiently address some limitations of linear modeling approaches, such as nonlinear relationships that are not adequately modeled by common transformations of variables. The group of procedures that includes GLMs and data mining techniques are often referred to as predictive models by insurance actuaries. In this paper we will show how data mining techniques and GLMs can be combined to take advantage of the strengths of each approach. In addition, we will present a common problem that arises in the modeling of count data: excess zeros. That is, sometimes, when actual instances of zero counts are compared to the theoretical values under the Poisson assumptions, there are significantly more zeros than the fitted distribution predicts. In the insurance context, this is believed to be due to the underreporting of small claims (Yip and Yau, 2005).

One of the symptoms of zero-inflated distributions is overdispersion. That is, under the Poisson assumption; the variance of the distribution is equal to its mean. Table 1.1 presents some automobile insurance count data from Yip and Yau that will be used throughout this paper to illustrate techniques and concepts. For this data the variance exceeds the mean. When the variance exceeds the mean, the situation is referred to as overdispersion, and a number of approaches are used to address it. One approach is to use a negative binomial model rather than a Poisson, as the variance of the negative binomial distribution exceeds the mean.

Table 1.1
Example of Overdispersion

K	Count	P(X=x)
0	1,706	0.607
1	351	0.125
2	408	0.145
3	268	0.095
4	74	0.026
5	5	0.002
Total	2,812	
Mean	0.815	
Variance	1.364	

Figure 1.1 displays a comparison of actual and theoretical probabilities at each value of K (or the five-year frequency) for the auto data. Note the actual data contains more zeros and fewer ones than predicted by the Poisson.

Figure 1.1
Actual Frequencies vs. Poisson Theoretical Frequencies

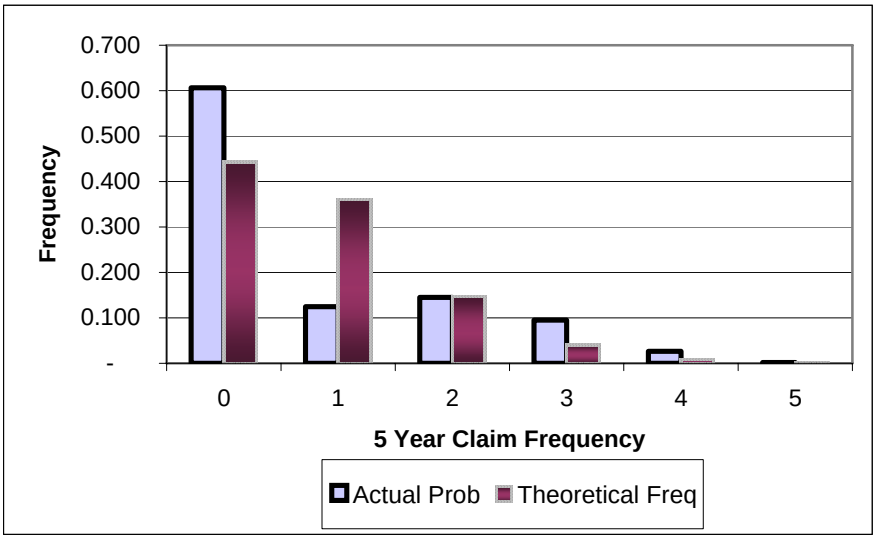


Table 1.2 displays the average claim frequency for the six car type categories in the data. The table indicates that some of the types such as Pickup and Van have similar frequencies. Might we be able to combine some of these categories and reduce the number of parameters in a regression model that uses categorical predictors? What procedures will facilitate efficiently combining of

categories that are not statistically different with respect to their effect on the dependent variable?

Table 1.2

Type of Car	Avg #Claims (Past 5 Years)
Panel Truck	0.9133
Pickup	0.8262
Sedan	0.6674
Sports Car	0.9296
SUV	0.8092
Van	0.8449
Total	0.8006

1.1 Research Context

As can be seen in some of the early literature on the subject (Bailey and Simon, 1959; Simon, 1962), the Poisson distribution has long been used in actuarial science as a stochastic model for claim count data. The negative binomial distribution is a key alternative when the variance of the count data exceeds the mean (Simon, 1962). Both distributions are members of the exponential family of distributions and have become popular for modeling frequency data in predictive modeling applications. Thus, the Poisson and negative binomial can be used within the GLM framework to fit regression models to insurance/claim frequency data.

Anderson et al. (2005) mention the problem of overdispersion that frequently occurs when using the Poisson distribution. Their suggested remedy follows that of the classic reference by McCulloch and Nelder (1989). The classical approach to overdispersion involves estimating an additional scale parameter for the Poisson distribution. This scale parameter has no effect on the estimated coefficients of the independent variables used in the regression model but does affect tests of significance for the variables. Ismail and Jemain (2007) extend the classical treatment of overdispersion using generalized Poisson and negative binomial models.

In Hilbe's recent book (Hilbe, 2007) points out that excess variability in Poisson regression can be due to a number of additional factors not remedied by using an overdispersion parameter or the negative binomial distribution including:

- Missing independent variables
- Interactions not included in the model

- Excess zeros

Yip and Yau (2005) illustrate how to apply zero-inflated Poisson (ZIP) and zero-inflated negative binomial (ZINB) models to claims data, when overdispersion exists and excess zeros are indicated. They also present another alternative, hurdle models, to approximate distributions with excess zeros. Jackman (2007) describes functions implemented in the statistical software R that can be used to implement ZIP, ZINB, and hurdle models. In this paper we will extend the work of these authors by combining ZIP, ZINB, and hurdle models with data mining procedures that efficiently search for significant terms in the data and reduce the dimensionality of categorical variables by clustering together categories of categorical dependent variables.

1.2 Objective

The paper attempts to improve the application of GLM procedures to claim prediction in property casualty insurance.

In this paper we will:

- Illustrate the problem of excess zeros in claim count data and then show how to remedy it with zero-adjusted mixture models
- Show how GLM models for count data can be combined with traditional data mining approaches to produce more robust models
- Apply the procedures to an insurance database as an illustration

1.3 Outline

The remainder of the paper proceeds as follows. Section 2 will present the problems of excess zeros in count data and show how to address it with zero-inflated models. In Section 3 we show how to augment GLM models with traditional data mining approaches to efficiently model nonlinear relationships and reduce the number of parameters contributed by categorical variables. In Section 4 we present overall conclusions. We have provided code in SAS for implementing some of the models in Appendices but numerous statistical tools contain the technology for implementing the models in this paper. Additional Code using R will be made available on the CAS's Web Site.

2. ZERO-INFLATED AND HURDLE MODELS

2.1 The Data

We will illustrate many of our key concepts using the auto data from Yip and Yau (2001). Yip and Yau supplied a frequency table of personal automobile claims that we use to illustrate Univariate distribution fitting methods. An additional database from Yip and Yau of personal automobile policy level information contains approximately 10,000 records and is used to illustrate multivariate regression models. Table 2.1 displays the variables in the data. The first variable on the list, claim frequency, is used as a dependent variable in the GLM, ZIP, and hybrid models. All other variables when used are used as predictor variables.

Table 2.1

Variables in Automobile Database

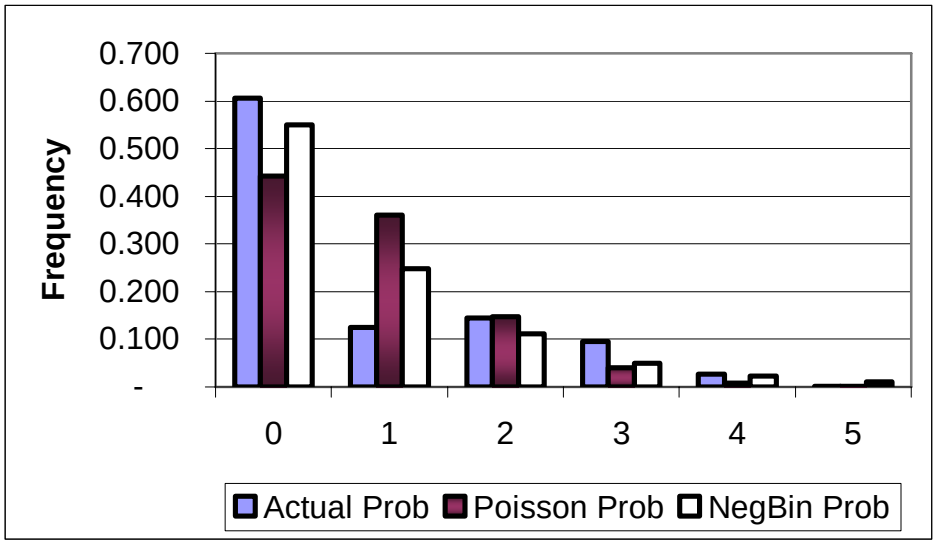
Variable	*Description*
CLM_FREQ	No. of claims in 5 years
AGE	Policyholder age
BLUEBOOK	Blue book value of car
CAR_TYPE	Type of car: sedan, SUV, etc.
CAR_USE	Private or Commercial use
CLM_DATE	Accident Date
DENSITY	Population Density (rural, urban)
GENDER	Gender
HOME_VALUE	House value
HOMEKIDS	No. of children at home
INCOME	Policyholder income
JOBCLASS	Job category
KIDSDRIVE	No. of children that drive
MARRIED	Marital status
MAX_EDUC	Highest education
MVR_PTS	Motor Vehicle Points
NPOLICY	Number of policies
PARENT1	Single Parent?
PLCYDATE	Policy Inception Data
RETAINED	Number of years policy renewed
REVOKED	Licensed revoked?
SAMEHOME	How many years in current house?
TRAVTIME	Travel time to work
YOJ	Years on current job

Before fitting a conditional model of claim frequency using the predictor variables in the auto data, we first investigate the distribution of marginal claims (displayed in Table 1.1). Figure 1.1 presented a comparison of actual and fitted Poisson claim frequencies for this data and indicated that the actual number of zero claims exceeds those that would be expected if the data were Poisson

distributed. A negative binomial distribution was fit next. A larger number of zeros (as well as larger frequencies) could be expected under a negative binomial model.

Figure 2.1

Comparison of Actual, Poisson, and Negative Binomial Frequencies



From Figure 2.1 it is apparent that the negative binomial distribution approximates the data better than the Poisson distribution. However, the actual data compared to the negative binomial shows an excess probability of zero claims and a significantly lower probability at a count of one.

2.2.1. Introduction to Zero-Inflated and Hurdle Probability Distributions

An alternative probability distribution when "excess" zeros appear to be present is the zero-inflated Poisson. The zero-inflated Poisson assumes the observed claim volumes are the result of a two-part process 1) a process that generates "structural zeros" and 2) a process that generates random claim counts. In insurance the "structural zeros" may be due to underreporting of small claims. Especially when claims are near or less than the policy deductible, a policyholder may not report the claim because 1) there may be no expected payment under the policy and 2) the policyholder may wish to avoid premium increases under an experience rating or merit rating system. The ZIP distribution is a mixture of exponential family distributions. Under the zero-

inflated Poisson, the structural zeros are assumed to follow a Bernoulli process with parameter φ, denoting the probability of a zero and the random counts are assumed to follow a Poisson with parameter λ, the mean of the distribution. The distribution of the zero-inflated Poisson is:

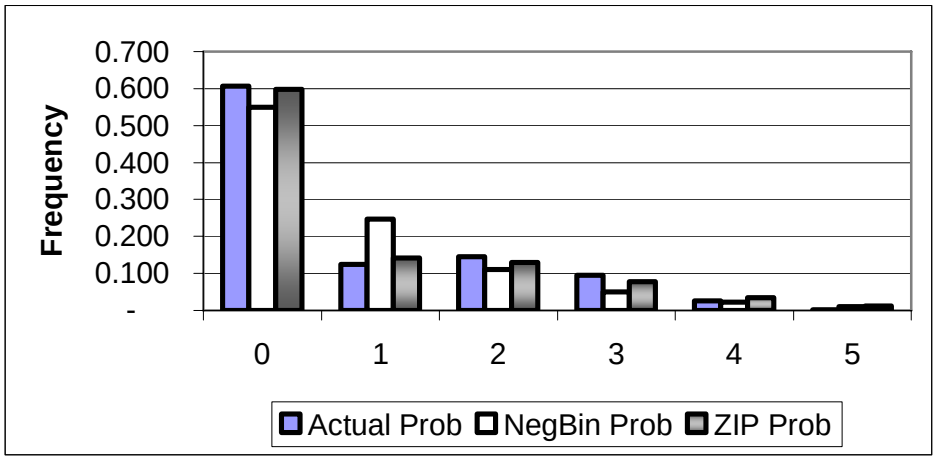
$$\begin{cases} \varphi + (1-\varphi)e^{-\lambda} & x=0 \\ (1-\varphi)\dfrac{\lambda^x}{x!}e^{-\lambda} & x>0 \end{cases} \tag{2.1}$$

The theoretical mean of the ZIP model is $\varphi + (1-\varphi)\lambda$. The variance is $(1-\varphi)\lambda(1+\varphi\lambda)$.

The parameters of the Poisson and negative binomial distribution can be estimated from the sample mean (Poisson) and the sample mean and variance (negative binomial). However, a numerical optimization procedure must be used to estimate the parameters of zero-inflated models. A description of the specific procedure we implemented in Excel is provided in Appendix G.

The parameters fit with Excel solver are displayed in Appendix G, Table G-2. The table indicates that on average, 54% of the records have structural zeros. For the remaining policyholders, the mean claim frequency over a five-year period is approximately 1.9. Figure 2.3 compares the negative binomial to the zero-inflated Poisson. The ZIP model appears to provide a better fit to the data.

Figure 2.3
Actual, Negative Binomial, and Zero-Inflated Poisson Frequencies



The Chi-Squared test can be used to test whether the ZIP model is a significantly better fit to the data than the negative binomial or Poisson models. The Chi-Squared statistic is:

$$\chi_{k-1}^2 = \frac{(Observed - Fitted)^2}{Fitted} \tag{2.3}$$

The Chi-Squared statistic compares the observed and fitted claim counts. It has degrees of freedom equal to k-1, where k is the number of categories (here equal to six).

Table 2.2
Chi Squared Statistic for Poisson, Negative Binomial and ZIP Models

	Chi-Squared
Model	**Statistic**
Poisson	**935.3**
Negative Binomial	**351.9**
ZIP	**60.2**

Note that the critical value for the Chi-Squared statistic at the 5% level is about 11, so that all three fitted models would be deemed significantly different from the data by this statistic.[1]

[1] It should be noted that a well-known limitation of the Chi-Square statistic is that it is very conservative when comparing actual to fitted distributions. That is, it is common for the distribution to be significantly different from the actual empirical distribution according to this measure.

However, it can be seen that the ZIP model provides a much better fit to the data.

Another mixed probability distribution related to the ZIP model is the zero-inflated negative binomial (ZINB) model. The ZINB is a mixture of a Bernoulli variable (for the structural zeros) and a negative binomial for the random counts. The distribution's formula is:

$$\begin{cases} \varphi + (1-\varphi)NB(0,r,p), k = 0 \\ (1-\varphi)NB(k,r,p), k > 0 \end{cases} \qquad (2.4)$$

$$NB(k,r,p) = \binom{k+r-1}{k} p^k (1-p)^r$$

The mean of the negative binomial is $r(1-p)/p$. The variance is $r(1-p)/p^2$. As with the ZIP model, the ZINB model can be fit in Microsoft Excel. Table G.3 in Appendix G shows the values of the estimated parameters.

In the example, the estimated parameters for ϕ is zero, indicating that the single negative binomial model is a better fit than the ZINB mixed model. Note the chi-square statistic for this model (397) was higher than that of the negative binomial fitted using the first two moments of the data.

A model related to the zero-inflated models is the hurdle model. The hurdle models assume two processes: 1) a process that generates no claim or at least one claim and 2) a process that generates the number of claims given that at least one claim occurs. A Bernoulli process is used to model the occurrence/nonoccurrence of a claim while a truncated Poisson or negative binomial is used to model positive claim counts. The formula for the hurdle Poisson model is shown in (2.5) and the fitted parameters are shown in Table G.4 of Appendix G. For this data the hurdle Poisson does not fit the data as well as the ZIP model, as it has a larger weighted squared deviation and its Chi-Square statistic of 97 is larger than that of the ZIP model.

$$\begin{cases} \varphi, k = 0 \\ \dfrac{1-\varphi}{1-e^{-\lambda}} \dfrac{\lambda^k}{k!} e^{-\lambda}, k > 0 \end{cases} \qquad (2.5)^2$$

[2] The mean of the hurdle Poisson is λ/ (1-exp(-λ)). The variance of the hurdle Poisson is λ/ (1-exp(-λ)) (1- λ exp(-λ))/ (1-exp(-λ)).

A negative binomial hurdle model was also fit to the data, but as with the ZINB model, the fitted model contained no Bernoulli parameter.

2.2.1.a Zero-Adjusted Models for Other Data Sets

Since the Yip and Yau data in our illustrations were used in their paper advocating the use of ZIP and ZINB models, one is not surprised when a zero-adjusted mixed model fits the data better than single count distribution models. In order to explore the broader applicability of zero-adjusted models, several other sample datasets were tested to determine if the ZIP or ZINB provided a better fit than simpler models:

- The Bailey and Simon credibility study (Bailey and Simon, 1959) used the experience from 1957 and 1958 for Canadian Private Passenger automobile exposure excluding Saskatchewan. The data is shown in Table 1 of their paper. This data is reorganized and displayed in Table F.1 of Appendix F. The data displayed was aggregated to the class level. For this data the negative binomial is a much better fit than the Poisson (illustrating the need to test for the negative binomial as an alternative to the Poisson), as well as the ZIP model. The ZINB, however, fits the data better than the negative binomial but the difference is not of the same magnitude as that between the negative binomial and Poisson. For this data, under the Poisson and ZIP assumptions observations are expected to be much closer to the distribution's mean value, while many of the actual observations are far from the mean, causing a very high chi-square values under Poisson and ZIP assumptions.
- Zero-inflated count data are also found in non-insurance applications. Five different datasets from various non-insurance analyses are displayed in Appendix F. Most of the examples tested displayed a very large variation in the goodness of fit. This wide variation indicates it may be prudent to test a number of possible alternatives before selecting a distribution to incorporate into a predictive model.
    - Hospital visit data from Deb and Trevedi (1997). The data contain the number of visits and hospital stays for a sample of United States residents aged 66 and over. For this data the ZINB was the best fit and the Poisson was a very poor fit.

- Doctor office visit data from Deb and Trevedi (1997). For this data the negative binomial was the best fit and the Poisson and ZIP were very poor fits.
- Patents data from Wang, Cockburn and Puterman (1998). The data contain the number of patents for a sample of pharmaceutical and biomedical companies. For this data the ZINB was the best fit and the Poisson and ZIP were very poor fits.
- Apple tree root cultivar[3] count data from Ridout and Demetrio (1998). For each cultivar, the number of roots produced during different experimental protocols was tabulated. For this data the ZINB was the best fit.

2.2.2. Poisson, Negative Binomial, ZIP, ZINB, and Hurdle Models with SAS

For simplicity of exposition, we have shown how to fit univariate zero-inflated and hurdle models in Microsoft Excel. However, nonlinear curve-fitting applications are typically performed in statistical or mathematical programming languages such as SAS, MATLAB, and R. For certain other distributions, specifically those that are members of the exponential family of distributions, a generalized linear model (GLM) can be use to fit the parameters of the distribution. For example an intercept-only GLM model with a Poisson distribution and log link can be used to estimate Poisson parameters. While this is a trivial example because the Poisson parameter equals its mean, it illustrates how common statistical software can be used to parameterize probability distributions. The model fit is:

(2.5) $Y = a + e$, where e is a random error term.

That is, a GLM procedure is used to fit a model that only has an intercept term, but no independent variables. For the Poisson, the intercept will equal the Poisson parameter. See Appendix B for an example of SAS code that can be used to fit the Poisson parameters.

For more complicated probability distributions such as zero-adjusted distributions, the analyst will want to use an approach that solves for parameters, given a function of the parameters to optimize. For instance in Appendix G the distance between an actual and fitted distribution is minimized when estimating the parameters of distributions using the Excel solver. It is common in distribution fitting to maximize the log of the likelihood function. For many common claim count

[3] A cultivar (short for cultivated variety) is a cultivated plant with unique characteristics that separate from other similar cultivated plants.

distributions, the log-likelihood function is readily specified, either from first principles or from one of the many references on probability distributions (Hogg and Klugman, 1982). In Appendix A we present the PDF and log-likelihood function for the Poisson. Once a likelihood function has been specified, an optimization procedure is used to solve for the distribution's parameters. For common one and two parameter distributions, it is often unnecessary to specify a likelihood function, as these functions are prepackaged in statistical fitting software.

For more complex models, many software packages offer the user a procedure that fits nonlinear mixed models using a nonlinear fitting procedure. This is appropriate for the zero-adjusted models, which do not have a closed-form solution for the parameters, but such procedures can often be used to fit more familiar distributions (i.e., Poisson, logNormal) as well (ignoring any "mixed" model structure). Appendix A presents an example using SAS code to generate fitted distributions and predicted probabilities.

Figures 2.1 and 2.2 suggest that the actual claim data contain excess zeros compared to those expected under both the Poisson and negative binomial distribution approximations. Prior to fitting a zero-inflated distribution, we can formally test for zero inflation. Van den Broek (1995) provides a score test for zero inflation relative to a Poisson distribution. The statistic is based on a comparison of actual zeros to those predicted by the model:

$$S = \frac{\left\{\sum_{i=1}^{n}(I(x_i = 0) - p_{0i}) / p_{0i}\right\}^2}{\sum_{i=1}^{n}(I(x_i = 0) - p_{0i}) / p_{0i} - n\bar{x}} \tag{2.6}$$

In formula (2.6) S is the score, $I(x_i=0)$ is an indicator function that is 1 if a given observation equals zero, and 0 otherwise. Denoting the probability, p_{0i} does so under the assumed distribution (typically Poisson) of a zero observation for observation i. Note that the probability is allowed to vary by observation. The score is assumed to follow a chi-squared distribution with one degree of freedom. Appendix C presents sample code that can be used to apply the score test. As seen in Appendix C, the score for our automobile count data was an 869, which is significant at the 0.001 level.

As the score statistic supports the possibility of a zero-inflated distribution, we proceed with fitting zero-inflated distribution using statistical software. Appendix D presents an example of

fitting a zero-inflated distribution using a nonlinear mixed models procedure.

As discussed in section 2.1.1, in the presence of excess zeros, a hurdle model rather than a zero-inflated model may be more appropriate. Hurdle (Mullahy,1986) or two-part (Heilbron, 1994) models are so-called because the likelihood function is constructed to be separable, that is, the zero/positive component is typically handled with a logistic or Probit model, whereas the model for positive counts can include or exclude zeros. The count portion of the hurdle model may be Poisson, negative binomial, or other count model. Appendix D presents SAS procedures that can be used to fit these hurdle models.

If zeros are excluded from the count portion of the model, then the positive portion can be modeled via a zero-truncated Poisson, for example. (The formula was given earlier in equation 2.5). Additional applications of truncated count models include Grogger and Carson (1991), Shaw (1988), and Winkelmann and Zimmerman (1995). Alternatives to the truncated Poisson include subtracting one from the dependent count variable. This has been described as a shifted or positive Poisson distribution (Shaw, 1988). Johnson and Kotz (1969) refer to this as a displaced Poisson distribution.

2.3 Regression Models

In this section, the zero-inflated and hurdle models are generalized to regression applications. We will use the 10,000 record Yip and Yau automobile insurance dataset to develop a model to predict claim frequency. This section will show how to augment the Poisson and negative binomial models commonly used for count predictions with zero-inflated and hurdle capabilities.

We first review the basic assumptions of generalized linear models. See Anderson et al. (2005) for a more complete introduction to GLMs.

A generalized linear model is denoted: $Y = \eta + e = \mathbf{x'\beta + e}$.

It has the following components:

- a random component, denoted e
- a linear relationship between a dependent variable and its predictors. The estimate or expected value of the prediction is denoted η.
- $\eta = a + b_1 x_1 + b_2 x_{2+...} b_n x_n$
- a link function captures the form of the relationship between the dependent variable and the

regression expected value. Two common link functions used when applying GLMs to ratemaking are:

- the identity link $\mu=\eta$
- the log link $\mu=\exp(\eta)$ or $\eta=\log(\mu)$.

Under the log link, each predictor variable's impact on the estimate is multiplicative. That is: $Y = A\exp(b_1 x_1)\exp(b_2 x_2)...\exp(b_n x_n)$. In ratemaking applications it is common for the classification variables to raise or lower a rate by a percentage. Hence, the log link is intuitive for the ratemaking models being presented in this paper.

Another common link function is the logit link: $\eta=\log(p/(1-p))$, where p denotes a probability between zero and one and $p/(1-p)$ is the odds ratio or the odds of observing the target variable. The logit link is commonly used with a Bernoulli (binary) dependent variable.

In claim frequency modeling, it is common for the random component of the GLM to be the Poisson or negative binomial distribution. The Poisson and under certain assumptions, the negative binomial (i.e., when the scale parameter is known) are members of the exponential family of distributions that also includes the normal and gamma. The zero-inflated and hurdle models generalize the GLM to include mixture models. For instance, the ZIP model is a mixture of two distributions from the exponential family: the Bernoulli and the Poisson. The hurdle Poisson model is also a mixture of a Bernoulli and a Poisson random variable, but with the hurdle model, the Poisson is a truncated Poisson that models only positive claim counts and the zeros are modeled exclusively with the Bernoulli distribution.

This paper's first predictive modeling illustration will use four variables to predict claim frequency. The four variables are car use, marital status, density, and gender. Each of the predictor variables is categorical. Thus the model is:

(2.7)1. $Y = f(\text{car use, marital status, density, gender}) + e.$

Where Y denotes the dependent variable, number of claims reported within a five-year period. In the Poisson and negative binomial regressions, the log link will be used.

In this section, classical GLM count regression models are compared to zero-inflated and hurdle alternatives. As discussed in Section 1, overdispersion in count models is commonly handled by

fitting an over-dispersed Poisson, which allows the variance to exceed the mean by a constant factor. We also present results for a geometric as well as a negative binomial model as the negative binomial becomes a geometric when the size parameter r is 1. Poisson and negative binomial regressions will be compared to ZIP, ZINB, and hurdle Poisson and hurdle negative binomial models. Under the zero-inflated and hurdle model there are two components denoted Y and Z:

$Y = f$(car use, marital status, density, gender) + e, Z = f(car use, marital status, density, gender) + e.

Thus, the predictor variables are used both to estimate the Bernoulli parameter p (the Z component) and are also used to estimate the Poisson expected claim count (the Y component). It is likely that the different variables will have a different importance in each component of the model. A nonlinear mixed models procedure can be used to estimate the parameters of the ZIP model. When using nonlinear mixed models procedures (or any other nonlinear optimization software) it is typically necessary to specify the log-likelihood function. For the ZIP regression the log-likelihood (denoted *ll*) is straightforward:

$$\begin{aligned} &\text{if } (Y=0)\ ll = \log(p_0 + (1-p_0)\exp(-\lambda)) \\ &\text{if } (Y>0)\ ll = \log(1-p_0) + Y * \log(\lambda) - \lambda - \log(x!) \end{aligned} \tag{2.8}$$

Appendix E presents code for fitting these models. In the particular example in Appendix E, the Bernoulli parameter p enters the function as a constant; that is, it is the same for every record, regardless of the value of the predictor variables, while the Poisson parameter is estimated from the regression function. It is straightforward to add a regression function for the Bernoulli parameter. To assess the goodness of fit of the models we compute the negative log-likelihoods (actually -2* the log-likelihood). In Table 2.3 the log-likelihood statistics from the different model fits are presented. It can be seen that the ZIP fits the data best while the simple Poisson regression provides the worst fit. Moreover, there is a significant improvement in fit when moving from the Poisson the ZIP.

The results indicate that the model fit to our sample auto claim counts was improved by using a zero-adjusted model. In the next section, we will compare and contrast a GLM and a zero-adjusted model with models augmented using hybrid techniques that employ a decision tree method to preprocess data. To keep the kinds of models to a manageable number we will only use the simple

Poisson and ZIP (the best performing model in Table 2.3) models in the next section.

Table 2.3

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

3. CHAID HYBRID MODELS

3.1 The CHAID method

The term "data mining" is loosely refers to a number of very different methods that apply computationally intensive nonparametric procedures, typically to large databases with many potential predictor variables and many records. Among the common data mining techniques used for prediction are neural networks and tree models. Trees fit a model by recursively partitioning the data into two or more groups, where data for each partition are more homogenous than the pre-partitioned data. The different groups are statistically determined to have significantly different values for the dependent variable. In the most common tree method, Classification and Regression Trees (C&RT), the data is split into two groups, one with a high average value for the dependent variable and the other group with a lower average value on the dependent variable. Each partition of the data in a tree model is referred to as a node.

The CHAID tree method is one of the oldest tree-based data mining methods and one of the earliest to appear in the casualty actuarial literature. The method was applied to classification ratemaking by Fish et al. (1990) following the passage of Proposition 103 in California.[4] Unlike C&RT, CHAID can partition data into more than two groups. CHAID is an acronym for chi-squared automatic interaction detection. As the name implies, CHAID relies heavily on the chi-squared statistic (Formula 2.3 in section 2) to partition data. In classical statistics the chi-squared statistic is typically used to assess whether discrete categorical variables are independent or whether a relationship exists between the variables (Faraway, 2006).

[4] Proposition 103 constrained how variables could be used in automobile ratemaking.

One of the data preparation steps that is often applied prior to fitting of predictive models is cardinality reduction. Cardinality reduction refers to reduction of the number of categories in nominal and ordinal variables (Refaat, 2007). The CHAID procedure is a procedure that can be used to preprocess categorical variables and to group like categories of the independent variables together. A problem with nominal and ordinal variables with many categories is that some of the categories are sparsely populated and some of the categories are very similar with respect to their effect on the dependent variable. Inclusion of all the levels of a categorical variable can lead to overfit/overparameterized models that fit parameters to noise rather than legitimate patterns in the data. Using the chi-squared statistic, categories that are not significantly different with respect to their effect on a dependant variable can be combined and the total number of categories reduced.

For instance, the categorical variable density from the automobile database has four levels or categories: highly urban, urban, rural, and highly rural. Suppose the analyst is interested in knowing whether a relationship exists between population density and the likelihood of having at least one claim. Let the likelihood of having a claim be denoted by a binary categorical indicator variable that is 1 if the policyholder has had at least 1 claim and 0 otherwise. Table 3.1 displays a crosstabulation of density and the indicator variable based on data from the automobile database. The bottom section of the table shows that urban and highly urban policyholders have a significantly higher frequency of claims than do rural and highly rural policyholders. The chi-squared statistic can be used to test whether this apparent relationship is significant.

Table 3.1

Crosstabulation of Population Density vs. Binary Claim Indicator

Home/Work Area * Claim Indicator Crosstabulation			
	Claim Indicator		
	No Claim	Claims	Total
Home/Work: Highly Rural	4,52	56	508
Home/Work: Highly Urban	1,732	1,867	3,599
Home/Work: Rural	1,369	196	1,565
Home/Work: Urban	2,740	1,891	4,631
Total	6,293	4,010	10,303

Percent of Policies With Claims			
	Claim Indicator		
	No Claim	Claims	Total
Home/Work: Highly Rural	89%	11%	100%
Home/Work: Highly Urban	48%	52%	100%
Home/Work: Rural	87%	13%	100%
Home/Work: Urban	59%	41%	100%
Total	61%	39%	

The chi-squared statistic requires both an observed and expected record count for each of the cells in the crosstabulation. An expected count can be computed by applying the marginal proportions shown at the bottom of Table 3.1 (61% no claim, 29% at least one claim) to the total number of policyholders in each density category. This is shown in Table 3.2. For instance, the expected number of highly rural drivers with no claims is 310.3 (0.89*508). The expected count is then used in the computation of the chi-squared statistic, shown also in Table 3.2. This statistic has degrees of freedom equal to the number (*c*-1)*(r- 1) (here 6) where *c* denotes the number of columns and *r* denotes the number of rows. Its value as shown at the bottom of Table 3.2, 886, is significant at (less than) the .1% level, suggesting a relationship between density and propensity for an automobile claim.

Table 3.2 Expected Count & Chi-Squared Statistic

Expected Count				
		Claim Indicator		Total
		No Claim	Claims	
Home/Work	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: $(O-E)^2/E$			
		Claim Indicator	
		No Claim	Claims
Home/Work	Highly Rural	64.70	101.60
	Highly Urban	98.90	155.20
	Rural	178.50	280.20
	Urban	2.80	4.40
			886.20

Suppose the claims are sorted in ascending order by proportion of policies with a claim. This is shown in Table 3.3. The table suggests that some of the categories of the density variable may not be significantly different from each other and therefore could be combined. For instance, the highly rural and rural categories at positive claim proportions of 11% and 13%, respectively, could perhaps be combined into a "rural" category, if the difference (in likelihood of having a claim) is not significant.

Table 3.3

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%

Table 3.4 displays the calculation of the chi-squared statistic, including the calculation of expected counts, for the highly rural and rural categories. The chi-squared statistic of 0.81 (see bottom row of Table 3.4) is not significant, indicating the two categories can be combined.

Table 3.4
Comparison of Rural and Highly Rural Categories Using Chi-Squared Statistic

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	

The chi-squared statistic can be computed for all other pairs of combinations (actually it only makes sense to compare pairs of categories that are contiguous in a sorted table such as Table 3.3). Once the chi-squared statistic has been computed for the pair-wise comparisons, the two categories with the lowest chi-squared values can be combined, provided the chi-square statistic is not significant.[5] In this example, the rural and highly rural categories have the lowest chi-squared statistics, so they are combined, resulting in three density groupings.[6] Table 3.5 shows the new table that is created when the categories are combined. Using the new crosstabulation, the chi-squared

[5] It is common to use the 5% level as the threshold for significance, though other levels can be chosen. Thus categories where the significance levels below the threshold can be combined. If the chi-squared statistic is significant, the two categories should not be combined, as the null hypothesis that there is no difference between the categories in their effect on the dependent variable is rejected.

[6] The chi-squared for all other comparisons was more than 99.0, which is significant at the 5% level.

statistic can be recomputed for the new table and the categories with the lowest chi-squared statistic can be combined. The recursive process of combining categories continues until no more significant differences between the categories can be found.

Table 3.5

Crosstabulation after Combining Two Categories

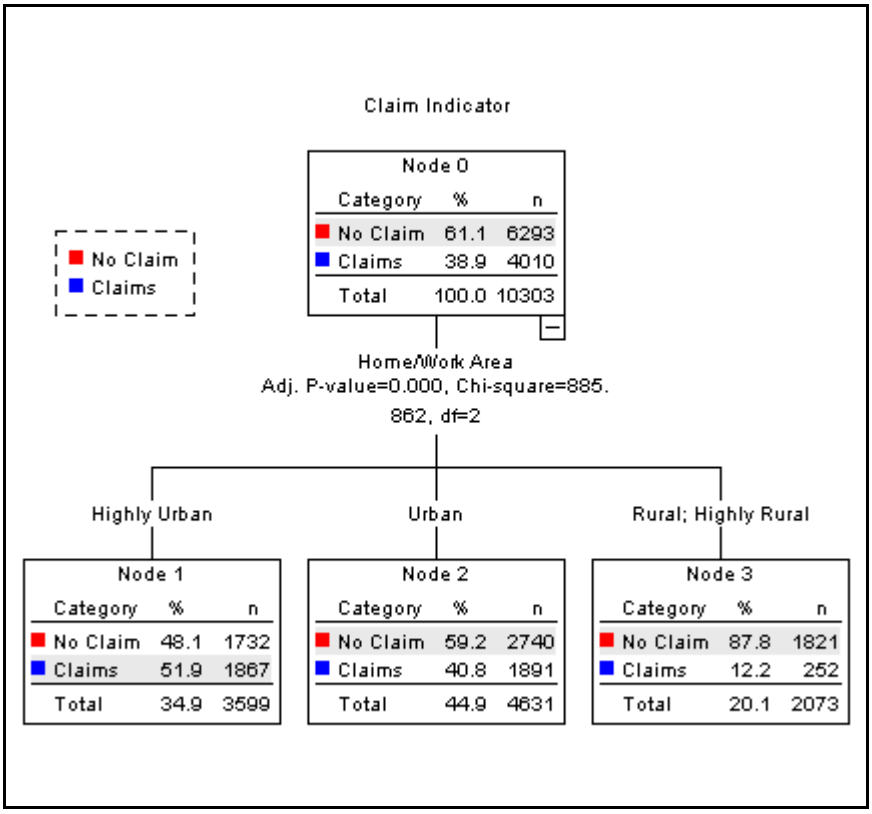
Area * Claim Indicator Crosstabulation			
	Claim Indicator		
	No Claim	Claims	Total
Rural	1,821	252	2,073
Urban	2,740	1,891	4,631
Highly Urban	1,732	1,867	3,599
Total	6,293	4,010	10,303
Percent of Policies With A Claim			
	Claim Indicator		
	No Claim	Claims	Total
Rural	88%	12%	508
Urban	59%	41%	3,599
Highly Urban	48%	52%	1,565

The results of the partitioning of the variables can be displayed graphically in a tree diagram. The tree diagram for the car density example is shown in Figure 3.1. The top box or "node" is a "parent" node. It displays the overall claim indicator statistics for all records before any partitioning occurs. Below the parent node are the "child" nodes resulting from the partitioning of the density variable using CHAID.[7] The nodes in this layer are also "terminal" nodes, as there is no further partitioning of the data. The terminal nodes contain the model's final prediction, which is typically the overall proportion of target variable records in the node.

[7] The CHAID models used in this paper were fit with SPSS Classification Trees. We are not aware of either SAS Stat or R functions for CHAID.

Figure 3.1

Tree for Population Density (Independent Variable) and Claim Indicator (Dependent Variable)



By adding a second variable to the model, say car use, it is possible to add another layer to the tree, however. To create a tree with two layers of nodes, it is necessary to partition the data on a second variable, after the partitions on the first variable, (density), have been completed. An example of partitioning using two variables is shown in Figure 3.2. As can be seen from Figure 3.2, not all nodes from the first layer can be further partitioned. When two variables are included in the model, CHAID performs the following process:

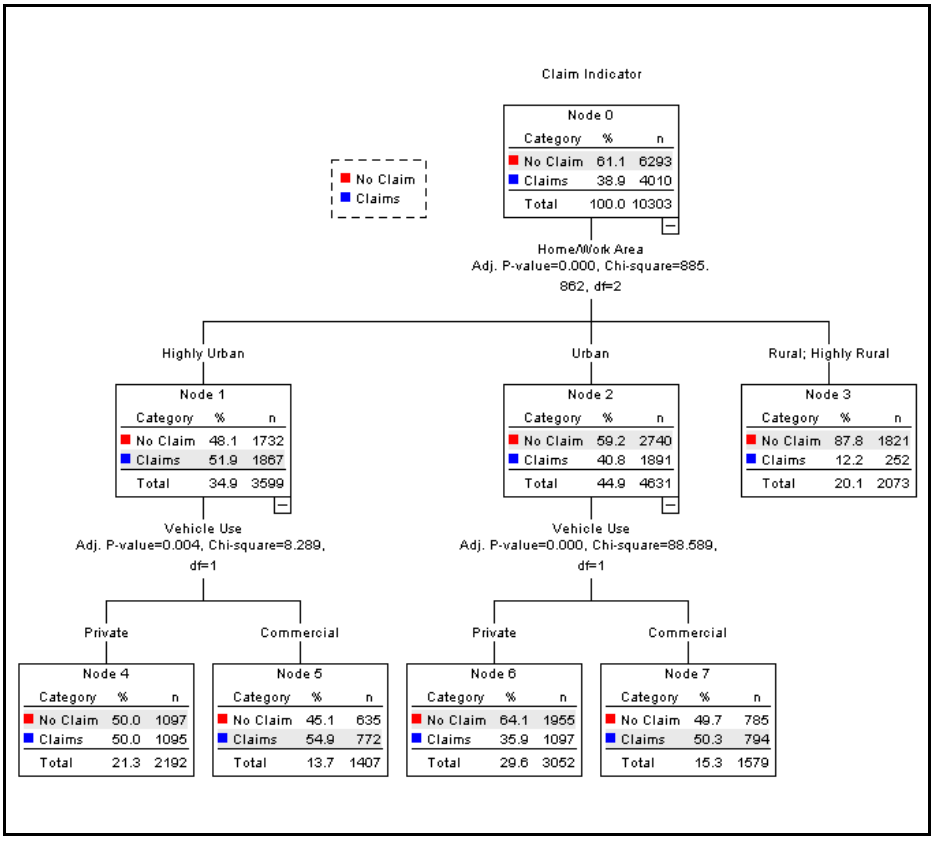
- Compute the beat partitioning of the data for the first variable and compute the chi-squared statistic for the partitioned data after categories that are not significantly different have been combined
- Compute the best partitioning of the data for the second variable and compute its chi-squared statistic. (In this very simple example, the car use variable has only two

categories, so no further combining of categories can be done)

- Select the variable that produces the highest chi-squared statistic to partition the data first
- Repeat the process for each of the nodes from the first partitioning. If none of the nodes can be further partitioned, stop.

Since, the focus of the current discussion is on the use of CHAID for cardinality reduction of categorical variables before fitting a GLM or other predictive model, further discussion of the CHAID for multivariable models is outside the scope of this paper. However, complete predictive models can be built using CHAID and other decision tree techniques.

Figure 3.2
CHAID Tree with Two Variables



It should be noted that the example of category reduction for the density variable is a relatively trivial one, as inspection of the statistical output from the fitted GLM, ZIP, and hurdle models could probably be used to reduce the number of categories. However, fast and computationally efficient procedures are needed for variables containing a large number of levels. Such variables occur frequently in insurance predictive models.

As a more realistic example, consider the car-type variable, which has six levels (Table 3.6). With six levels for a variable, there are hundreds of possible ways to combine categories.[8] In a typical automobile ratemaking database, there would likely be many more than six levels on a car-type variable. Figure 3.3 presents the CHAID tree that was fit using the car-type variable.

[8] The number of all possible combinations is $\sum_k \binom{x}{k}$, but when the categories are ordered based on the proportion of policies with claims, the number goes down.

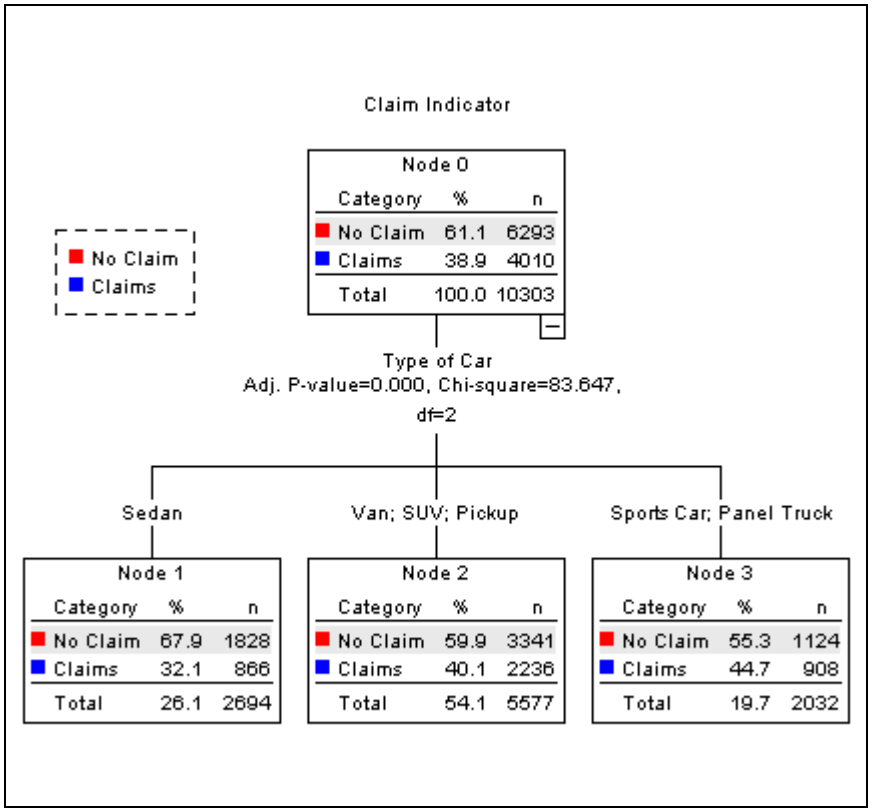
Table 3.6

Car Type Frequency Tabulation

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%

Figure 3.3

CHAID Model for Car-type Variable



From Figure 3.3, the number of groupings is reduced from six to three when CHAID is used to preprocess the car-type variable.

When the dependent variable in the model is numeric, rather than categorical, most CHAID

procedures use the *F*-statistic rather than the chi-squared statistic to partition data.

(3.1)
$$F = \frac{(RSS_1 - RSS_2)/(p_2 - p_1)}{RSS_2 / p_2}$$

RSS = residual sum of squares

p_1 = degrees of freedom for model 1

p_2 is degrees of freedom for model 2

When only two categories are compared, the *F*-test reduces to a *T* test.[9] Thus the categories can be compared using the *F* (or *T*) statistic and the categories that are not significantly different can be merged. The first two categories merged are the categories with the lowest *T* statistic.

Suppose, instead of using a binary categorical dependent variable, we treat claim frequency (number of claims in the past five years) as a numeric variable and use the *T* test to merge categories. Table 3.7 displays the mean claim frequency, along with standard deviations and confidence intervals for the density variable. It is clear that the rural and highly rural categories can be merged, as their claim frequencies are the same.

Table 3.7

Mean Five-Year Claim Frequency by Density

					95% Confidence Interval for Mean	
	N	Mean	Std. Deviation	Std. Error	Lower Bound	Upper Bound
Highly Rural	508	.24	.758	.034	.18	.31
Highly Urban	3,599	1.07	1.223	.020	1.03	1.11
Rural	1,565	.24	.707	.018	.21	.28
Urban	4,631	.84	1.171	.017	.81	.88
Total	10,303	.80	1.154	.011	.78	.82

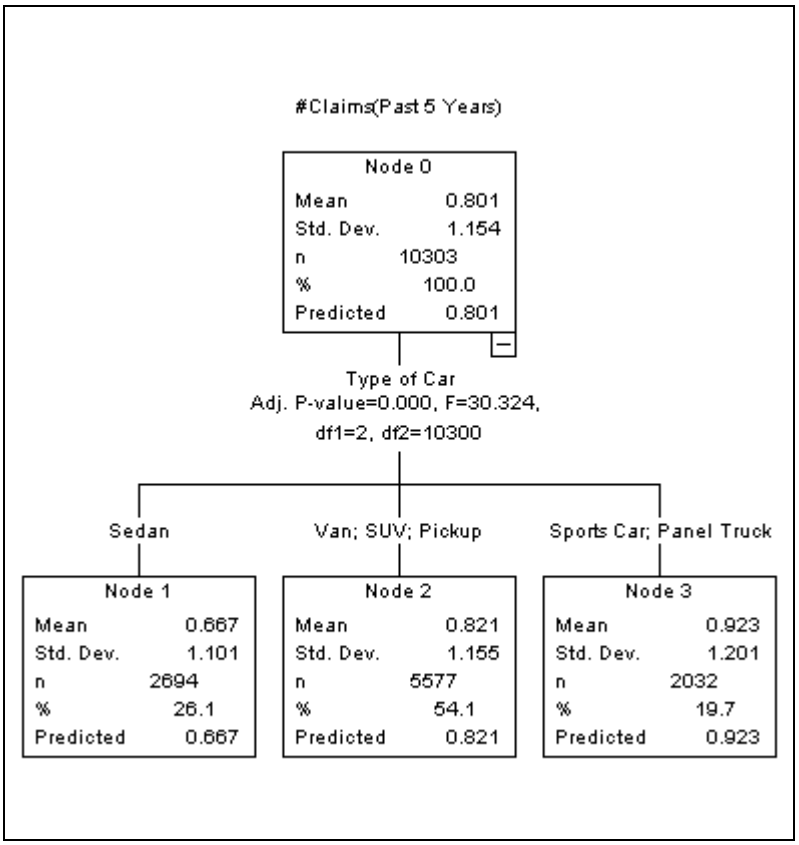
Figure 3.4 shows that if claim frequency is treated as a numeric variable, and is used to group the

[9] T= $(\bar{x}_1 - \bar{x}_2)/s_{\bar{x}}, \bar{x}$ is mean of group and $s_{\bar{x}}$ is sd of difference between means

categories of the car-type variable, the same grouping is created as for the binary claim indicator variable in Figure 3.3, which used the chi-squared statistic to partition data.

Figure 3.4

CHAID Tree for Car Type with Numeric Dependent Variable



A new categorical variable can be created using the results of the CHAID analysis. The new car-type variable has three rather than six categories. Two predictive models were than fitted, using the new variables 1) a Poisson regression and 2) a ZIP regression. As a measure of goodness of fit, we use the Akaike Information Criterion (AIC) statistic. This statistic penalizes the log of the likelihood function when degrees of freedom, i.e., additional parameters, are incorporated into the model. Each variable in the model adds to its degrees of freedom. A model with a categorical variable having six levels adds five degrees of freedom[10] to the model, while a variable having three levels adds only two degrees of freedom. The formula for the AIC is:

[10] One degree of freedom for each binary dummy variable created which is k-1, where k is the number of categories

(3.2) $$AIC = 2 * df - 2 * \log\ likeihood$$

From Table 3.8, the AIC statistic indicated a better fit for both the Poisson regression and the ZIP regression when the car-type variable has been preprocessed to reduce the number of categories.

Table 3.8

Akaike Information Criterion, Car Type and Grouped Car Type

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

The CHAID procedure can also be used to preprocess numeric variables. The relationship between continuous independent variables and a dependent variable is frequently nonlinear. One way to model the nonlinearities is to bin the numeric variables. When a variable is binned, ranges of the variable are grouped together and treated as a level of a categorical variable. Thus, claimant ages can be binned into 0 – 10, 11 – 20, etc. Tree procedures such as CHAID can be used to optimally bin numeric variables (Refaat, 2007). To illustrate how this can be done, the CHAID procedure will be used to bin the motor vehicle record (i.e., the number of points on the policyholder's record) variable from the automobile data. Table 3.9 displays a frequency distribution for the motor vehicle record variable. It can be seen that the number of points ranges from 0 to 13.

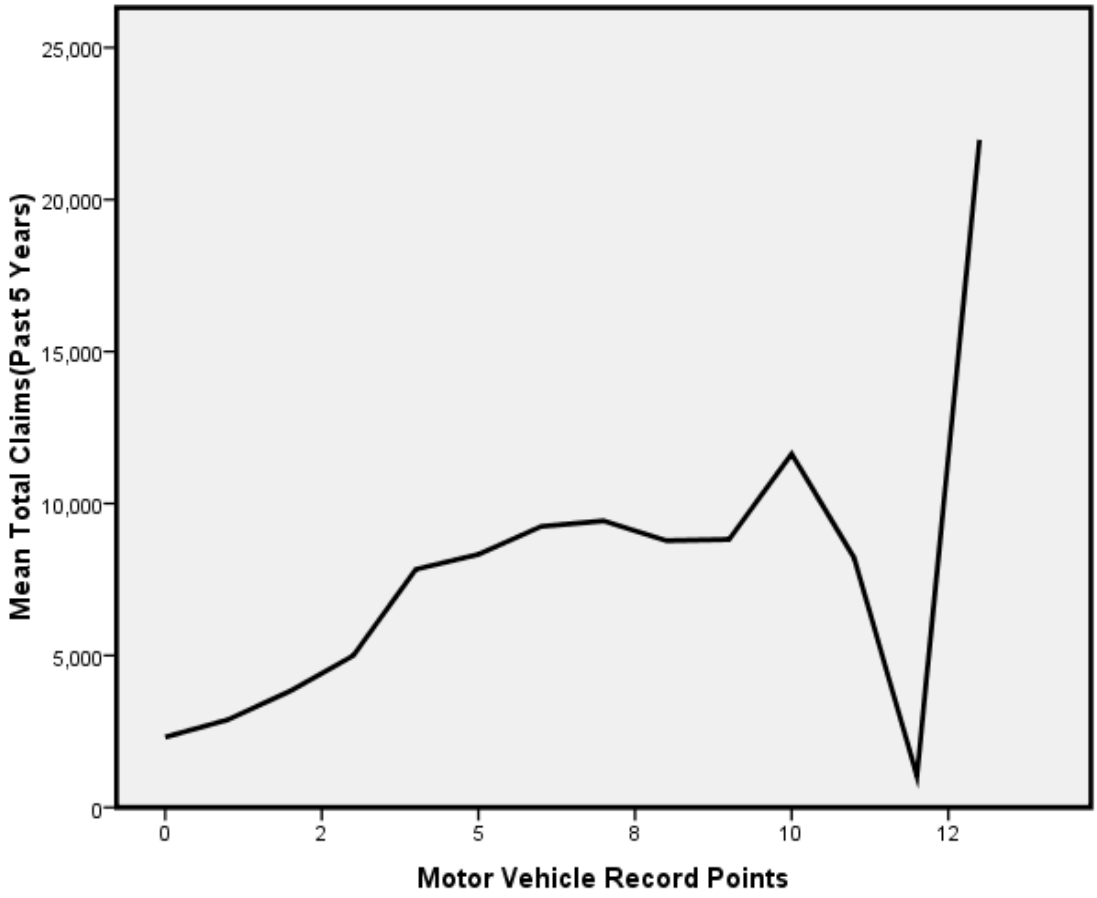
The distribution is a skewed distribution. That is, most of the values exceed the distribution's median. About 45% of policyholders have no points and 60% have one or fewer points. Figure 3.5, which displays the average frequency by motor vehicle record, indicates that the relationship between motor vehicle record and frequency is nonlinear. Claim frequency increases between zero and about five points and then (ignoring the inherent variability at high point values due to the sparseness of the data) appears to level off.

Table 3.9

Frequency Distribution for Motor Vehicle Record

Motor Vehicle Record Points			
Liscense Points	Frequency	Percent	Cumulative Percent
0	4,659	45.2	45.2
1	1,467	14.2	59.5
2	1,199	11.6	71.1
3	966	9.4	80.5
4	727	7.1	87.5
5	528	5.1	92.7
6	341	3.3	96
7	213	2.1	98
8	114	1.1	99.1
9	53	0.5	99.7
10	20	0.2	99.8
11	13	0.1	100
12	1	0	100
13	2	0	100
Total	10,303	100	

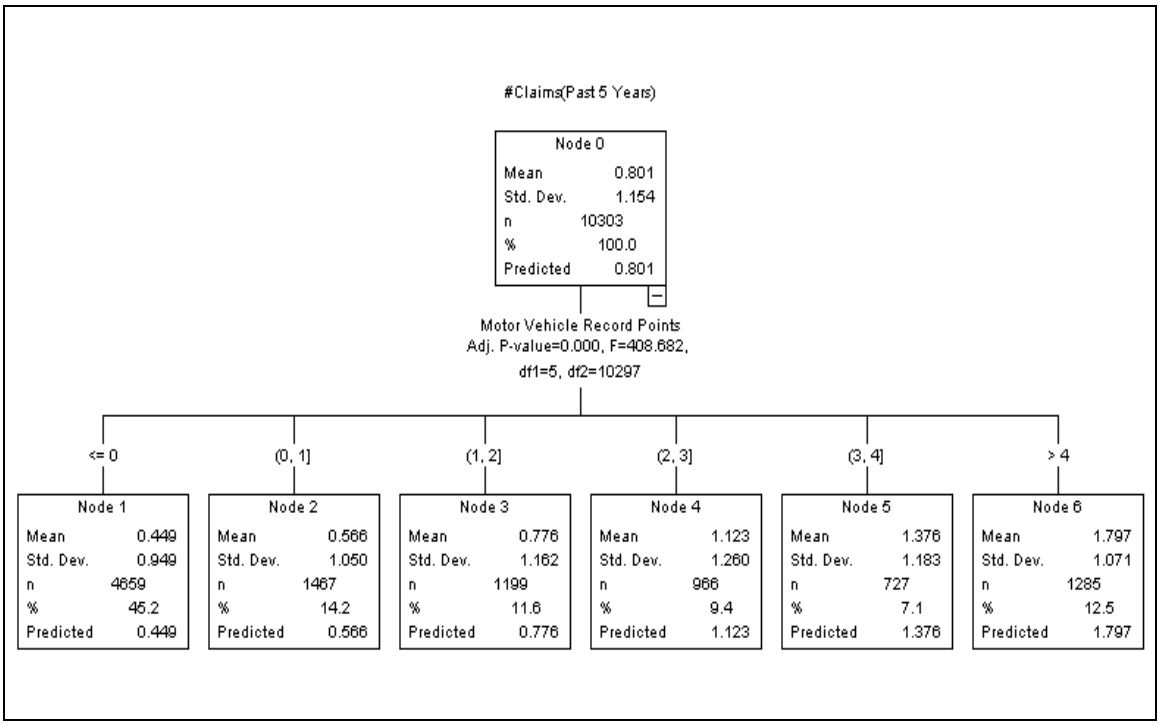
Figure 3.5
Average Claim Frequency by Motor Vehicle Record



How can the analyst best bin the motor vehicle record variable to approximate the relationship between motor vehicle points and claim frequency? One approach is to use the CHAID procedure to group together values of motor vehicle record with similar values for average claim frequency. Figure 3.6 displays the output of the CHAID procedure for motor vehicle record. Figure 3.6 indicates that each value from one through four is significantly different from other values and that it should stand alone as a bin. In predictive modeling, once the motor vehicle records have been binned, the new variable containing the binned categories can be used as a nominal variable in a regression. Alternatively, Figure 3.5 suggests that the relationship between motor vehicle record and claim frequency may be linear until about a value of 5 and then level off.

Figure 3.6

Tree Displaying Bin for Motor Vehicle Record



To test which treatment of the motor vehicle variable might work best, both a Poisson and ZIP regression were fit using the original variable, the variable capped at a value of 5 and the binned variable. For both the Poisson and the ZIP model, the binned variable performed better than the capped or original variable when AIC is used a goodness-of-fit measure. The lowest AIC was for the ZIP model with MVR binned.

Table 3.10

AIC for Original Variable, Capped Variable and Binned Variable

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

3.2 Results for Multi-Variable Model

To test the different methods a model that contained six variables (car use, gender, marital status, density, car type, and motor vehicle points) was fit. The number of categories for the density and car-type variables was reduced using CHAID. The motor vehicle record variable had two scenarios: MVR capped and MVR binned. The results for the Poisson regression and the ZIP regression are displayed in Table 3.11. Table 3.11 indicates that preprocessing improves the fit of the Poisson regression. The improvement was approximately the same whether motor vehicle record was capped or binned. On the other hand, the fit of the ZIP model declined when motor vehicle record was capped, but improved when it was binned. The AIC statistics in Table 3.11 also indicate that the ZIP model provides a significantly better fit than the Poisson model.

Table 3.11

AIC for Full Regression, Original Data and Preprocessed Data

Treatment of Variable	Poisson	ZIP
Original Variables	12,066	10,622
CHAID, MV Capped	12,009	10,676
CHAID, MV Binned	12,003	10,546

3.2.1 Out of Sample Goodness-of-Fit Measures

In predictive modeling, it is customary to test models on a sample of data that was set aside specifically for that purpose. The data used in this paper was split into two samples: a "training" sample used to fit the model's parameters and a "testing" sample used to test the fit of the model, once parameters have been estimated using the "training" sample.

In typical insurance databases, traditional measures of goodness of fit often perform poorly. For instance, the R^2 for the zip model applied to the test sample is 0.22, a number which, although low, is higher than what can be obtained in most data bases, likely because the frequencies in the data are based on five years of experience. In an automobile insurance data base where frequencies are based on annual experience, perhaps 90% of policyholders will not have experienced a claim, even though all policyholders have some positive expectation of a claim. Thus the actual value for most records

will be "0" while the predicted value will be greater than "0" resulting in an R^2 statistic that tends to be low. To provide a useful test, comparisons must be based on aggregates of the data that are sufficiently large that the mean frequency in a group will be greater than zero. One way to aggregate the data is to create groups based on the value of the model's predicted value. The predicted value is sometimes referred to as the model's "score." All records can be sorted based on their score. The test data can then be grouped into quantiles based on the model score. For instance, the data can be split into ten groups based on the model score assigned to each record. For each decile, the actual frequency from the data can be computed. A graph comparing the actual to the predicted values within each decile can be created and used to visually evaluate the fit.

Figure 3.7

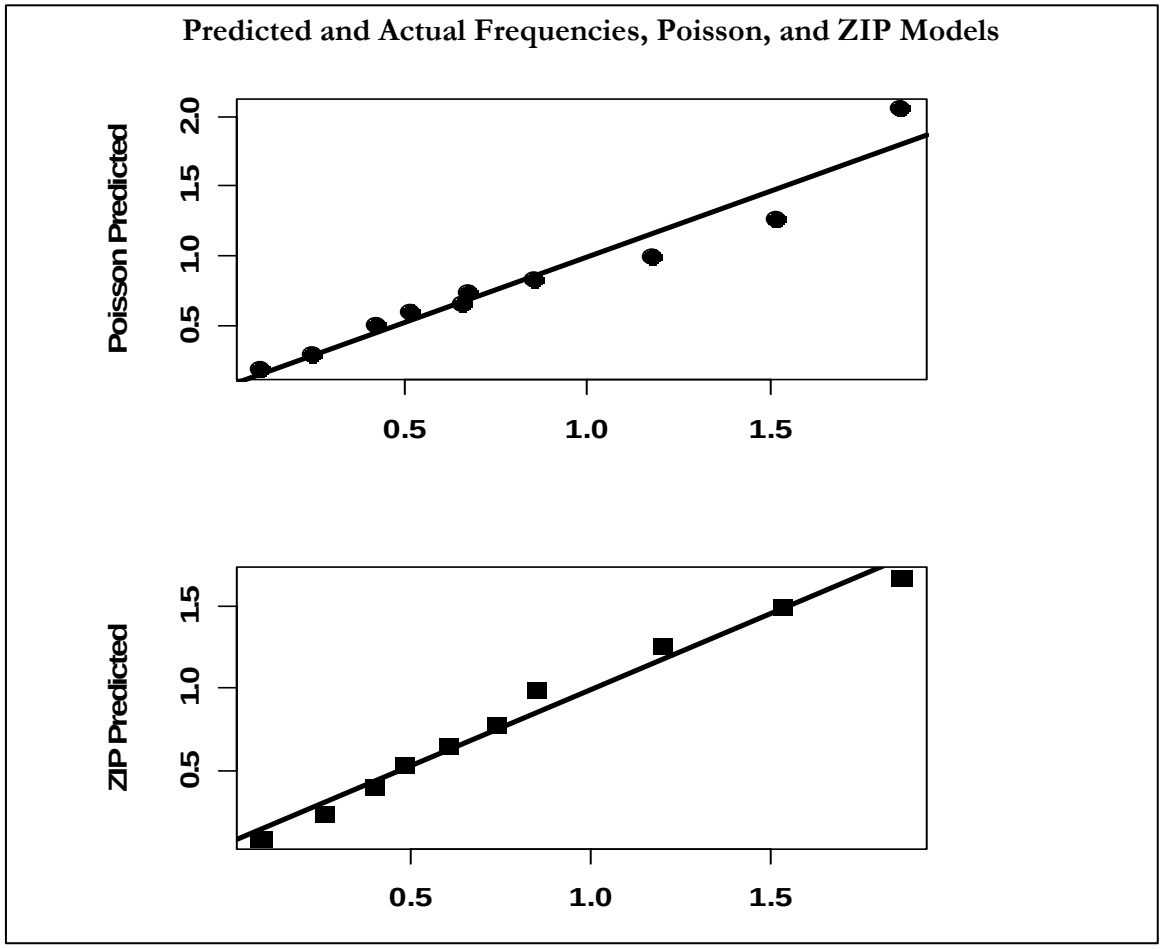


Figure 3.7 displays a comparison of actual and predicted frequencies for test data grouped by decile of the models score. A model with good predictive ability should be upward sloping; for each

increase in decile based on model score, the actual frequency should increase. Also, a high correlation between actual and predicted frequency indicates a good fit. In Figure 3.7 the best fitting line based on ordinary least squares regression is shown. A high correlation between actual and predicted values is indicated by a small scatter of points around the line. In table 3.12, the correlation coefficient of the six models on the test data is shown.

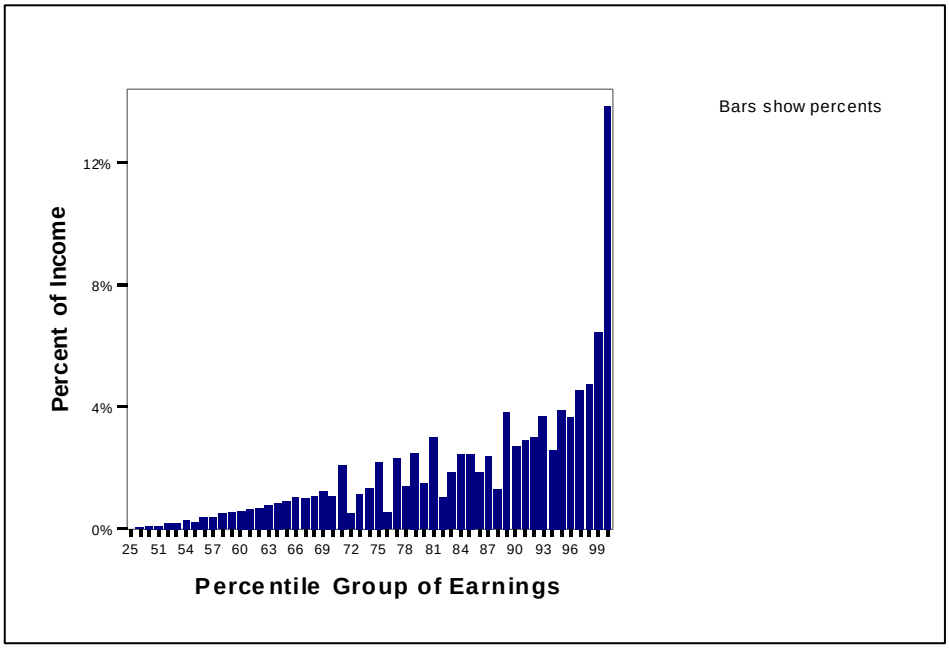
Table 3.12

Treatment of Variable	Poisson	ZIP
Original Variables	0.9720	0.9900
CHAID, MV Capped	0.9860	0.9900
CHAID, MV Binned	0.9810	0.9940

The correlations in Table 3.12 indicate that the ZIP models fit the out of sample data better than the Poisson models. It also indicates that preprocessing of variables with CHAID improves the fit of the Poisson regression models, but appears to have minimal effect on the ZIP models.

Meyers (2006) presented another curve that can be used to visualize the fit of models on out of sample data. The curve is based on the Lorenz curve. The Lorenz curve arose out of studies of income inequality by 19th and 20th century economists (Arnold, 1983). For example, Figure 3.8 displays a distribution of incomes from the 2000 census for the state of Pennsylvania. From this graph, it can be seen that earners in the highest percentiles earned a disproportionate share of incomes. The top 1% of individuals earned 13% of the state of Pennsylvania's income.

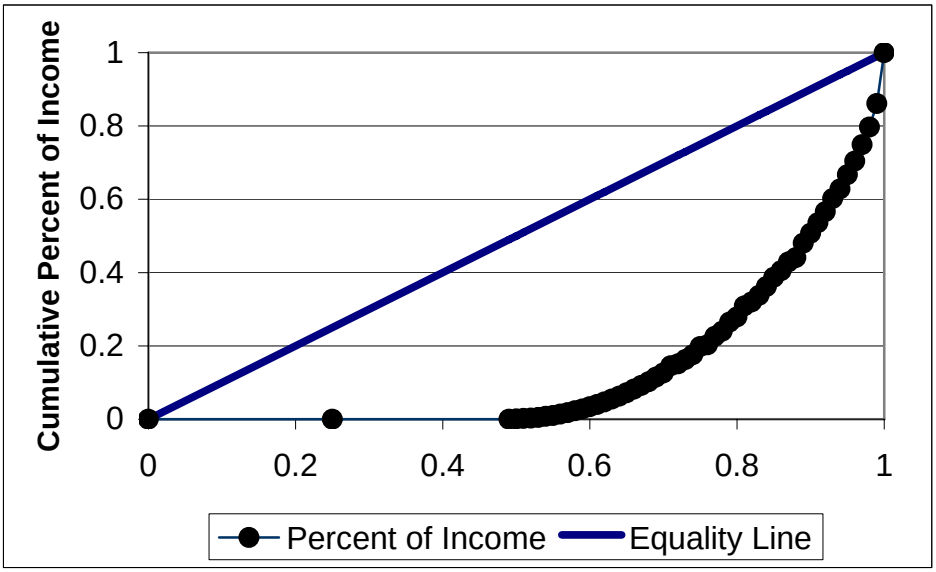
Figure 3.8: Income Distribution from 2000 Census



By cumulating the data from Figure 3.8, i.e., computing the cumulative percent of all income earned by a given percent of the population, a Lorenz curve can be created. This is shown in Figure 3.9.

Figure 3.9

Lorenz Curve for Income from 2000 Census



If income distribution were perfectly equal, incomes would be distributed according to the diagonal line. The area between this line and the curve for the income distribution is a measure of income known as the Gini Index. The greater the income inequality, the larger the Gini Index should be. A simple formula for this area based on the trapezoidal rule for numerical integration (Press et al., 1989) is:

$$A = \left(\frac{1}{2} F_1 + F_{2\ldots} + F_{n-1} + \frac{1}{2} F_n \right) * \Delta x, \tag{3.5}$$

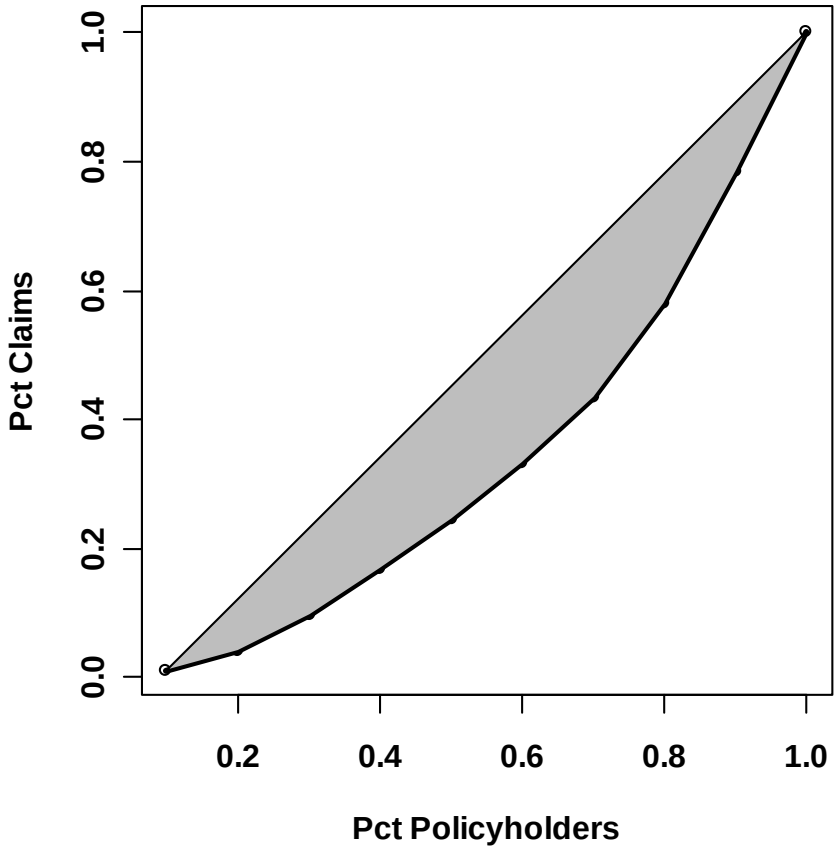
F_i is ith cumulative income

The statistic in (3.5) is also known as the Gini Index. It was introduced by Meyers (2006) as a general procedure for assessing the fit of models. A Lorenz curve can also be constructed from predictive models and the insurance data they are applied to. Figure 3.9 displays an approximation to the Lorenz curve based on the Poisson model. This approximation is based on only 10 groups or deciles of the data,[11] although often more intervals are used.

[11] The test was limited to deciles, as a model with categorical predictors that have only a few levels may have a limited number of possible values.

Figure 3.9

Lorenz Curve for Poisson Predictive Model



The data was also used to compute an approximation to the Gini Index. Table 3.13 presents the approximation Gini Index for the six models.

Table 3.13
Gini Index for Models

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

The out of sample tests in Table 3.13 indicate that the ZIP model fits better than the Poisson. For the Poisson model, it also indicates that preprocessing the variables using CHAID to construct a capped MVR variable improves the fit of the Poisson model but not of the ZIP model.

4. CONCLUSIONS

In this paper, alternatives to the Poisson and negative binomial distributions for count regressions were presented. One alternative makes use of mixed zero-adjusted (zero-inflated and hurdle) distributions. These are mixture models composed of two distinct probability distributions, thus the resulting distribution is not a member of the exponential family of distributions. The alternative provided a significantly better fit to a database of automobile insurance claims than did the Poisson and negative binomial models. Moreover, many other authors (Yip and Yau, 2005; Heilbron, 1994) use zero-inflated and hurdle models to better approximate data than simple Poisson and negative binomial models. In our day-to-day property/casualty insurance modeling, we have found that zero-inflated and hurdle models frequently fit the data better than Poisson and negative binomial models. We have found this to be the case across a number of different lines, including homeowners, personal automobile, and workers compensation. The phenomenon of excess zeros is also commonly encountered in non-insurance applications such as quality control (Lambert, 1992) and biostatistics (Ridout et al., 1998). We tested a small selection of non-insurance databases and zero-adjusted distributions provided a better fit to some of the data. Thus it seems appropriate to test for excess zeros using a test such as Van den Broek's score test. See Appendix C for more information on this test. If excess zeros are indicated, either a zero-inflated or hurdle model is likely to provide a better model than a classical Poisson or negative binomial regression. The testing displayed wide variation between the goodness of fit of the different distributions assessed, suggesting that it is prudent to test several alternative distributions before fitting a model.

A limitation of many GLMs that incorporate categorical variables is over-parameterization. This occurs when more categories are included than are needed. When categories that are statistically equivalent are combined, the over-parameterization is eliminated. In this paper a relative quick and efficient procedure for reducing the cardinality of nominal variables was presented. The procedure in this paper used the CHAID decision tree procedure to statistically determine the appropriate way to combine categories. This paper provided an example where application of the CHAID procedure

to group categories of categorical variables improved the fit of the model. Typical predictive modeling applications databases contain a number of categorical variables with many levels. For instance, there may be 100 or more different types of vehicles in a vehicle-type variable, and many are sparsely populated. Since the categorical variables in our data had only a relatively small number of categories, the benefit of preprocessing categorical data was illustrated but could not be fully exploited.

Another limitation of GLMs with numeric predictor variables is that the relationship between the predictor and dependent variable may be nonlinear. The CHAID technique can be used to preprocess numeric variables to approximate the nonlinear relationship.

Supplementary Material

Excel spreadsheets, R and SAS Code will be available on the CAS Web Site.

Appendix A
The Poisson Distribution and the Use of Nonlinear Mixed Modeling Procedures to Fit Parameters

We begin this illustration by using SAS Proc NLMIXED to estimate the parameters of the Poisson. We will derive the log-likelihood of the Poisson from its PDF, to illustrate its use in Proc NLMIXED.

The Poisson PDF is

$$(Y = y \mid \mu) = \frac{e^{-\lambda}\lambda^{y}}{y!} \text{ where } y = 0, 1, 2, 3, \ldots$$

with:

$$mean = E(Y \mid X = x_i) = \lambda$$

$$\text{var} = \sigma^2 = Var(Y \mid X = x) = \lambda$$

It is clear from the above formulas that the mean of the Poisson equals its parameter, lambda.

Differentiating the PDF with respect to our mean parameter, lambda; the log-likelihood:

$$ll=-\lambda+y*\log(\lambda)-lgamma(y+1)\ .$$

Below we illustrate the use of this function in a SAS procedure that is used to estimate the parameter of the Poisson. We also illustrate how to directly fit the Poisson, without specifying a likelihood function. Proc NLMIXED is designed to estimate the parameters of nonlinear mixed models. A mixed model arises when some of the independent variables in a model are themselves random realizations from a distribution rather than fixed quantities (see Venables and Ripley, 2002; Faraway, 2006). A discussion of mixed models is beyond the scope of this paper; however, knowledge of how to specify random effects is unnecessary when using Proc NLMIXED to fit common probability distributions such as the Poisson. Below is the SAS code used for the fit:

```
proc sql;
      select max(clm_freq) into :y_max from claims2;
quit;
```

```
%put *** y_max=&y_max.;

%macro estimate;
      %do i = 0 %to &y_max.;
            estimate "p(&i.)" pdf('poisson',&i., lambda);
      %end;
%mend;
proc nlmixed data=claims2;
      parms eta=-0.2;
      lambda = exp(eta);
      y = clm_freq;
      model y ~ poisson(lambda);
/* or /
      loglike = -lambda + y*log(lambda) - lgamma(y + 1);
      model y ~ general(loglike);
/* or /
      pdf = (exp(-lambda)*lambda**y)/fact(y);
      loglike=log(pdf);
      model y ~ general(loglike);
      estimate 'lambda' lambda;
      %estimate;
      predict lambda out=predpoi(keep=clm_freq pred);
      title 'Poisson model via Proc NLMIXED';
run;
title;
```

Poisson model via Proc NLMIXED

The NLMIXED Procedure

Specifications	
Data Set	WORK.CLAIMS2
Dependent Variable	y
Distribution for Dependent Variable	Poisson
Optimization Technique	Dual Quasi-Newton
Integration Method	None

Dimensions	
Observations Used	2812
Observations Not Used	0
Total Observations	2812
Parameters	1

Parameters	
eta	*NegLogLike*
-0.2	3782.77991

Iteration History					
Iter	*Calls*	*NegLogLike*	*Diff*	*MaxGrad*	*Slope*
1	4	3782.75728	0.022624	1.211746	-51.3544
2	5	3782.75696	0.00032	0.002708	-0.00064
3	6	3782.75696	1.728E-9	7.161E-7	-3.2E-9

NOTE: GCONV convergence criterion satisfied.

Fit Statistics	
-2 Log Likelihood	7565.5
AIC (smaller is better)	7567.5
AICC (smaller is better)	7567.5
BIC (smaller is better)	7573.5

Parameter Estimates									
Parameter	*Estimate*	*Standard Error*	*DF*	*t Value*	*Pr > |t|*	*Alpha*	*Lower*	*Upper*	*Gradient*
eta	-0.2045	0.02089	2812	-9.79	<.0001	0.05	-0.2454	-0.1635	7.161E-7

Additional Estimates								
Label	*Estimate*	*Standard Error*	*DF*	*t Value*	*Pr > |t|*	*Alpha*	*Lower*	*Upper*
lambda	0.8151	0.01703	2812	47.87	<.0001	0.05	0.7817	0.8485

SAS Proc NLMIXED also makes it very simple to fit a negative binomial distribution to the sample data. Again, there several ways to specify the distribution. The ability to code with programming statements within Proc NLMIXED is very flexible. One can use the internal specification for negative binomial, specify the negative binomial PDF, take the log and use the model general option, or directly specify the log-likelihood one wishes to solve for it directly. Beside the model fit, we can also ask for some additional statistics such as contrast testing for whether our Negbin

dispersion is significantly different from zero (Poisson), as well as the estimated variance and predicted probabilities for each count.

```
%macro estimate;
      %do i = 0 %to &y_max.;
      estimate "p(&i.)"  (gamma(&i. + k)/(gamma(&i. + 1)*gamma(k)))*
                         (((1/k)*mu)**&i.)/(1 + (1/k)*mu)**(&i. + (k));
      %end;
%mend;

proc nlmixed data=claims2;
      parms  b_0=-.2 k=1.4;
      eta = b_0;
      mu = exp(eta);
      y = clm_freq;
/* specify the full log-likelihood */
/*    loglike = (lgamma(y + (1/k)) - lgamma(y + 1) - lgamma(1/k) + */
/*                         y*log(k*mu) - (y + (1/k))*log(1 + k*mu)); */
/*    model y ~ general(loglike);*/
/* or, use the internal negbin(n,p) representation */
      p = exp(-eta)/(1 + exp(-eta));
      model y ~ negbin(1/k,p);
      predict mu out=out2(keep=clm_freq pred);
      contrast 'k = 0' k - 0;
      estimate 'exp(b_0)' exp(b_0);
      estimate 'mean' mu;
      estimate 'k' k;
      estimate 'variance' mu + k*mu**2;
      %estimate;
      title 'Negative Binomial model via Proc NLMIXED';
run;
```

Specifications	
Data Set	WORK.CLAIMS2
Dependent Variable	y
Distribution for Dependent Variable	Poisson
Optimization Technique	Dual Quasi-Newton
Integration Method	None

Dimensions	
Observations Used	2812
Observations Not Used	0
Total Observations	2812
Parameters	1

Parameters	
eta	*NegLogLike*
-0.2	3782.77991

Iteration History					
Iter	*Calls*	*NegLogLike*	*Diff*	*MaxGrad*	*Slope*
1	4	3782.75728	0.022624	1.211746	-51.3544
2	5	3782.75696	0.00032	0.002708	-0.00064
3	6	3782.75696	1.728E-9	7.161E-7	-3.2E-9

NOTE: GCONV convergence criterion satisfied.

Fit Statistics	
-2 Log Likelihood	7565.5
AIC (smaller is better)	7567.5
AICC (smaller is better)	7567.5
BIC (smaller is better)	7573.5

Parameter Estimates									
Parameter	*Estimate*	*Standard Error*	*DF*	*t Value*	*Pr > |t|*	*Alpha*	*Lower*	*Upper*	*Gradient*
eta	-0.2045	0.02089	2812	-9.79	<.0001	0.05	-0.2454	-0.1635	7.161E-7

Additional Estimates								
Label	*Estimate*	*Standard Error*	*DF*	*t Value*	*Pr > |t|*	*Alpha*	*Lower*	*Upper*
lambda	0.8151	0.01703	2812	47.87	<.0001	0.05	0.7817	0.8485

Negative Binomial model via Proc NLMIXED

Specifications	
Data Set	WORK.CLAIMS2
Dependent Variable	y
Distribution for Dependent Variable	Negative Binomial
Optimization Technique	Dual Quasi-Newton
Integration Method	None

Dimensions	
Observations Used	2812
Observations Not Used	0
Total Observations	2812
Parameters	2

Parameters		
b_0	*k*	*NegLogLike*
-0.2	1.4	3560.13868

Iteration History					
Iter	*Calls*	*NegLogLike*	*Diff*	*MaxGrad*	*Slope*
1	3	3502.65393	57.48475	37.59117	-2543.71
2	4	3501.12368	1.53025	11.94101	-2.31042
3	5	3500.97924	0.144435	1.374487	-0.31788
4	6	3500.97779	0.001458	0.050886	-0.003
5	7	3500.97778	1.233E-6	0.00302	-2.59E-6

NOTE: GCONV convergence criterion satisfied.

Fit Statistics	
-2 Log Likelihood	7002.0
AIC (smaller is better)	7006.0
AICC (smaller is better)	7006.0
BIC (smaller is better)	7017.8

Parameter Estimates									
Parameter	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper	Gradient
b_0	0.1388	0.07768	2812	1.79	0.0741	0.05	-0.01354	0.2911	0.002728
k	1.4095	0.1006	2812	14.01	<.0001	0.05	1.2122	1.6068	-0.00302

Contrasts				
Label	Num DF	Den DF	F Value	Pr > F
k = 0	1	2812	196.23	<.0001
k = 1	1	2812	16.56	<.0001

Additional Estimates								
Label	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper
exp(b_0)	1.1489	0.08924	2812	12.87	<.0001	0.05	0.9739	1.3238
mean	1.1489	0.08924	2812	12.87	<.0001	0.05	0.9739	1.3238
k	1.4095	0.1006	2812	14.01	<.0001	0.05	1.2122	1.6068
variance	3.0092	0.5030	2812	5.98	<.0001	0.05	2.0229	3.9956

Appendix B
Count Distribution Parameter Estimation Using SAS Proc GENMOD

Below a Poisson distribution is fit to the data with Proc GENMOD, SAS Generalized Linear Model procedure. The estimate statement applies the inverse of the log link, eponentiating the intercept displaying the estimated mean as 0.82 along with a 95% confidence interval of (0.7824,0.8491). Note that the log likelihood reported in Proc GENMOD is not directly comparable to those reported in Proc NLMIXED or some other software as Proc GENMOD

drops the *y*-factorial component of the likelihood as this does not contribute to estimating the mean parameter and can cause numeric instabilities with high counts.

```
proc genmod data=claims2;
      model clm_freq = / link=log dist=Poisson noscale;
      estimate 'mean' intercept 1 / exp;
      title 'Poisson Distribution';
run;
```

SAS Proc GENMOD also makes it very simple to fit a negative binomial distribution to our sample data. Here we simply change the dist= option to our model statement.

```
proc genmod data=claims2;
      model clm_freq = / link=log dist=NegBin;
      estimate 'mean' intercept 1 / exp;
      title 'NegBin model';
run;
```

NegBin Model

The GENMOD Procedure

Model Information		
Data Set	WORK.CLAIMS2	
Distribution	Negative Binomial	
Link Function	Log	
Dependent Variable	CLM_FREQ	#Claims(Past 5 Years)

Number of Observations Read	2812
Number of Observations Used	2812

Parameter Information	
Parameter	*Effect*
Prm1	Intercept

Criteria For Assessing Goodness Of Fit			
Criterion	*DF*	*Value*	*Value/DF*
Deviance	2811	2549.6471	0.9070
Scaled Deviance	2811	2549.6471	0.9070
Pearson Chi-Square	2811	2190.0408	0.7791
Scaled Pearson X2	2811	2190.0408	0.7791
Log Likelihood		-2478.8688	

Analysis Of Parameter Estimates							
Parameter	*DF*	*Estimate*	*Standard Error*	*Wald 95% Confidence Limits*		*Chi-Square*	*Pr > ChiSq*
Intercept	1	-0.2045	0.0306	-0.2645	-0.1445	44.59	<.0001
Dispersion	1	1.4095	0.1006	1.2123	1.6067		

Note: The negative binomial dispersion parameter was estimated by maximum likelihood.

Contrast Estimate Results							
Label	*Estimate*	*Standard Error*	*Alpha*	*Confidence Limits*		*Chi-Square*	*Pr > ChiSq*
mean	-0.2045	0.0306	0.05	-0.2645	-0.1445	44.59	<.0001
Exp(mean)	0.8151	0.0250	0.05	0.7676	0.8655		

Appendix C

SAS Code for SCORE Test

```
/*************************************************************/
/* Van den Broek (1995) score test                           */
/* Van den Broek, Jan,
 A score test for zero inflation in a Poisson distribution,
 Biometrics, 1995, v51, n2, p738-743                         */
/*************************************************************/
proc sql;
      select sum(((clm_freq=0) - exp(-pred))/exp(-pred))**2 as num,
                  sum((1 - exp(-pred))/exp(-pred)) -
count(clm_freq)*mean(clm_freq) as denom,
                       count(clm_freq) as n, mean(clm_freq) as ybar,
```

```
            (sum(((clm_freq=0) - exp(-pred))/exp(-pred))**2) /
                  (sum((1 - exp(-pred))/exp(-pred)) -
count(clm_freq)*mean(clm_freq)) as score,
                        1 - probchi(calculated score, 1) as p format 8.6
                  into :num, :denom, :n, :ybar, :score, :p
                        from out2;
quit;
%put ****************************************************************;
%put *** Van den Broek - Score statistic for extra zeros;
%put *** num=&num, denom=&denom., n=&n., ybar=&ybar., score=&score., p=&p.;
%put ****************************************************************;
```

Van den Brock – Score Statistic for extra zeros

numerator	denom	n	ybar	score	p
1,086,713.00	1,249.30	2812	0.8151	869.9	0.000001

Appendix D

SAS Code for Zero-Inflated Models

This appendix shows how to fit a zero-inflated distribution in SAS Proc NLMIXED. (Also, an experimental procedure under the SAS/ETS product, Proc COUNTREG directly fits ZIP and ZINB models. See http://support.sas.com/kb/26/addl/fusion26161_3_countreg.pdf. Here, we add some options. The parameters statement allows us to specify starting values for our parameters to be estimated. The bounds statement allows us to constrain our zero-inflation factor to the logical range. Again, we utilized the flexibility of programming and the estimate statements to calculate several useful estimates such as the expected number of conditional ZIP mean and variance.

```
%macro estimate;
      %do i = 0 %to &y_max.;
       %if &i.=0 %then %do;
      estimate "p(&i.)" p_0 + (1 - p_0)*pdf('poisson',&i., lambda);
      %end;
            %else %do;
            estimate "p(&i.)" (1 - p_0)*pdf('poisson',&i., lambda);
      %end;
      %end;
%mend;
proc nlmixed data=claims3;
  parameters  p_0=0.57 bll_0=0.5;
  bounds 0<p_0<1;
  eta = bll_0;
  lambda = exp(eta);
  y = clm_freq;
  if y=0 then loglike = log(p_0 + (1 - p_0)*exp(-lambda));
     else loglike = log(1 - p_0) + y*log(lambda) - lambda - lgamma(y + 1);
  model y ~ general(loglike);
  contrast 'p_0 - 0' p_0 - 0;
  estimate "p_0" p_0;
  estimate "Expected zeros=exp(-lambda)" exp(-lambda);
  estimate 'Conditional Poisson Mean (lambda)' lambda;
  estimate 'ZIP Mean (1-p_0)*lambda' (1 - p_0)*lambda;
  estimate 'ZIP Var(1-p_0)*lambda*(1+lambda+(1-p_0)*lambda)'
                    (1 - p_0)*lambda*(1 + lambda + (1 - p_0)*lambda);
/*  estimate "Proportion of 'extra' zeros (theta)" theta; */
  estimate 'theta=p_0/(1-p_0)' p_0/(1 - p_0);
  %estimate;
  predict p_0 out=pred_zi(keep=pred);
  predict lambda out=pred(keep=clm_freq pred);
  predict (1 - p_0)*lambda out=out2(keep=clm_freq pred);

  title 'Zero-Inflated Poisson (ZIP) distribution';
run;
title;
```

Zero-Inflated Poisson (ZIP) Distribution

Specifications	
Data Set	WORK.CLAIMS3
Dependent Variable	y
Distribution for Dependent Variable	General
Optimization Technique	Dual Quasi-Newton
Integration Method	None

Dimensions	
Observations Used	2812
Observations Not Used	0
Total Observations	2812
Parameters	2

Parameters		
p_0	*bll_0*	*NegLogLike*
0.57	0.5	3360.90317

Fit Statistics	
-2 Log Likelihood	6695.2
AIC (smaller is better)	6699.2
AICC (smaller is better)	6699.2
BIC (smaller is better)	6711.1

Parameter Estimates

Parameter	*Estimate*	*Standard Error*	*DF*	*t Value*	*Pr > |t|*	*Alpha*	*Lower*	*Upper*	*Gradient*
p_0	0.5177	0.01231	2812	42.04	<.0001	0.05	0.4935	0.5418	4.247E-7
bll_0	0.5247	0.02658	2812	19.74	<.0001	0.05	0.4726	0.5768	6.041E-6

Contrasts				
Label	Num DF	Den DF	F Value	Pr > F
p_0 - 0	1	2812	1767.27	<.0001

Additional Estimates								
Label	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper
p_0	0.5177	0.01231	2812	42.04	<.0001	0.05	0.4935	0.5418
Expected zeros=exp(-lambda)	0.1845	0.008289	2812	22.26	<.0001	0.05	0.1683	0.2008
Conditional Poisson Mean (lambda)	1.6899	0.04492	2812	37.62	<.0001	0.05	1.6018	1.7780
ZIP Mean (1-p_0)*lambda	0.8151	0.02331	2812	34.96	<.0001	0.05	0.7694	0.8608
ZIP Var(1p_0)*lambda*(1+lambda+ (1-p_0)*lambda)	2.8568	0.1254	2812	22.79	<.0001	0.05	2.6110	3.1026
theta=p_0/(1-p_0)	1.0733	0.05293	2812	20.28	<.0001	0.05	0.9695	1.1771

Given the additional flexibility introduced with the zero-inflation parameter, The zero-inflated negative binomial (ZINB) distribution fit estimates a very small dispersion parameter, *k*.

```
proc nlmixed data=claims3;
      parms bp_0=.07 bll_0=0.52 k=0.000033;
      bounds k>0;
      eta_zip = bp_0;
      p0_zip = exp(eta_zip)/(1 + exp(eta_zip));
      eta_nb = bll_0;
      mean   = exp(eta_nb);
      y = clm_freq;
      p0 = p0_zip + (1 - p0_zip)*exp(-(y + (1/k))*log(1 + k*mean));
      p_else = (1 - p0_zip)*exp(lgamma(y + (1/k)) - lgamma(y + 1) –
            lgamma(1/k) + y*log(k*mean) - (y + (1/k))*log(1 + k*mean));
      if y=0 then loglike = log(p0);
          else    loglike = log(p_else);
    model y ~ general(loglike);
    estimate "Estimated proportion of 'extra' zeros (theta)" p0_zip;
    estimate 'Estimated Conditional Poisson Mean (Lambda)' mean;
    estimate 'Estimated Unconditional ZIP Mean' (1-p0_zip)*mean;
    estimate 'Estimated Unconditional ZIP Variance'
                                  (1-p0_zip)*mean*(1+p0_zip*mean);
    predict mean out = mean_hat;
    title 'Zero-inflated Negative Binomial ZINB Distribution';
run;
```

Zero-Inflated Negative Binomial ZINB Distribution

Specifications	
Data Set	WORK.CLAIMS3
Dependent Variable	y
Distribution for Dependent Variable	General
Optimization Technique	Dual Quasi-Newton
Integration Method	None

Dimensions	
Observations Used	2812
Observations Not Used	0
Total Observations	2812
Parameters	3

Parameters			
bp_0	*bll_0*	*k*	*NegLogLike*
0.07	0.52	0.000033	3347.62028

Fit Statistics	
-2 Log Likelihood	6695.2
AIC (smaller is better)	6701.2
AICC (smaller is better)	6701.2
BIC (smaller is better)	6719.0

Parameter Estimates

Parameter	*Estimate*	*Standard Error*	*DF*	*t Value*	*Pr > |t|*	*Alpha*	*Lower*	*Upper*	*Gradient*
bp_0	0.07089	0.04949	2812	1.43	0.1521	0.05	-0.02614	0.1679	0.083701
bll_0	0.5246	0.02698	2812	19.44	<.0001	0.05	0.4717	0.5775	-0.11562
k	1.187E-6	0.001040	2812	0.00	0.9991	0.05	-0.00204	0.002041	229.649

Additional Estimates								
Label	*Estimate*	*Standard Error*	*DF*	*t Value*	*Pr > |t|*	*Alpha*	*Lower*	*Upper*
Estimated proportion of 'extra' zeros (theta)	0.5177	0.01236	2812	41.90	<.0001	0.05	0.4935	0.5419
Estimated Conditional ZINB Mean (Lambda)	1.6898	0.04559	2812	37.06	<.0001	0.05	1.6004	1.7793
Estimated Unconditional ZINB Mean	0.8150	0.02340	2812	34.83	<.0001	0.05	0.7691	0.8609
Estimated Unconditional ZINB Variance	1.5280	0.05533	2812	27.61	<.0001	0.05	1.4195	1.6365

Appendix D

SAS Code for Hurdle Model

```
/*************************************/
/* fit Hurdle - Binomial for 0/1     */
/*************************************/
data claims4;
      set claims2;
      clm=(clm_freq>0);
run;
proc nlmixed data=claims4;
      parms b_o=-0.52;
      y = clm;
      eta = b_o;
      p1 = exp(eta)/(1 + exp(eta));
      model y ~ binary(p1);
      estimate 'phi' 1-1/(1 + exp(-b_o));
run;
/*************************************/
/* fit Hurdle - Truncated Poisson    */
/*************************************/
proc nlmixed data=claims3(where=(clm_freq>0));
  *   parms b_0=0.52;
      eta_lam = b_0;
    lambda = exp(eta_lam);
      y = clm_freq;
      prob = ((exp(-lambda)*(lambda**y))/fact(y))/(1 - exp(-lambda));
      loglike = log(prob);
      model y ~ general(loglike);
      estimate 'lambda' lambda;
      estimate 'conditional mean' lambda/(1 - exp(-lambda));
      estimate 'conditional var' (lambda/(1 - exp(-lambda)))*
                      (1 - (lambda*exp(-lambda))/(1 - exp(-lambda)));
      predict (lambda/(1 - exp(-lambda))) out=tpois_pred;
      title 'Count model for non-Zero Outcomes (Poisson)';
run;
title;
```

Appendix E

Code for Fitting Models

Poisson, negative binomial, ZIP, ZINB, and hurdle GLM regression models are fit with SAS. Below a Poisson regression model is fit to the data with Proc GENMOD. Main effects regressors are added to same setup as above: car use, marital status, area, and sex.

```
/*******************************************************/
/* fit Poisson regression model (including covariates) */
/*******************************************************/
proc genmod data=claims2;
      class car_use mstatus area sex;
      model clm_freq = car_use mstatus area lincome sex
                                / link=log dist=Poisson;
run;
```

The GENMOD Procedure

Model Information		
Data Set	WORK.CLAIMS2	
Distribution	Poisson	
Link Function	Log	
Dependent Variable	CLM_FREQ	#Claims(Past 5 Years)

Number of Observations Read	2812
Number of Observations Used	2812

Criteria For Assessing Goodness Of Fit			
Criterion	*DF*	*Value*	*Value/DF*
Deviance	2811	4736.2388	1.6849
Scaled Deviance	2811	4736.2388	1.6849
Pearson Chi-Square	2811	4706.1012	1.6742
Scaled Pearson X2	2811	4706.1012	1.6742
Log Likelihood		-2760.6479	

Analysis Of Parameter Estimates							
Parameter	*DF*	*Estimate*	*Standard Error*	*Wald 95% Confidence Limits*		*Chi-Square*	*Pr > ChiSq*
Intercept	1	-0.2045	0.0209	-0.2454	-0.1635	95.82	<.0001
Scale	0	1.0000	0.0000	1.0000	1.0000		

Note: The scale parameter was held fixed.

Model Information		
Data Set	WORK.CLAIMS2	
Distribution	Poisson	
Link Function	Log	
Dependent Variable	CLM_FREQ	#Claims(Past 5 Years)

Number of Observations Read	2812
Number of Observations Used	2812

Class Level Information		
Class	*Levels*	*Values*
CAR_USE	2	Commercial Private
mstatus	2	1. Yes 2. No
area	2	1. Highly Urban/ urban area 2. Highly Rural/ rural area
sex	2	1. M 2. F

Criteria For Assessing Goodness Of Fit			
Criterion	*DF*	*Value*	*Value/DF*
Deviance	2806	4312.6656	1.5369
Scaled Deviance	2806	4312.6656	1.5369
Pearson Chi-Square	2806	4636.5810	1.6524
Scaled Pearson X2	2806	4636.5810	1.6524
Log Likelihood		-2548.8614	

Analysis Of Parameter Estimates

Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept		1	-1.1801	0.1013	-1.3787	-0.9815	135.66	<.0001
CAR_USE	Commercial	1	0.2831	0.0449	0.1951	0.3711	39.73	<.0001
CAR_USE	Private	0	0.0000	0.0000	0.0000	0.0000	.	.
mstatus	1. Yes	1	-0.0961	0.0425	-0.1794	-0.0127	5.10	0.0239
mstatus	2. No	0	0.0000	0.0000	0.0000	0.0000	.	.
area	1. Highly Urban/ urban area	1	1.3631	0.0835	1.1994	1.5268	266.28	<.0001
area	2. Highly Rural/ rural area	0	0.0000	0.0000	0.0000	0.0000	.	.
lincome		1	-0.0206	0.0061	-0.0327	-0.0086	11.32	0.0008
sex	1. M	1	-0.1206	0.0441	-0.2070	-0.0343	7.49	0.0062
sex	2. F	0	0.0000	0.0000	0.0000	0.0000	.	.
Scale		0	1.0000	0.0000	1.0000	1.0000		

Note: The scale parameter was held fixed.

It is also fairly simple to add regressors to the linear predictor, eta, in Proc NLMIXED. One small complication is that Proc NLMIXED does not offer a class statement, therefore one has to either create desired indicator or dummy variables ahead of time, or as in the example below, use programming statements to create them "on-the-fly," The phrase inside each set of parentheses resolves to either true or false, zero, or one.

```
data claims3;
      set claims2;
      car_usen=0; if car_use='Commercial' then car_usen=1;
      mstatusn=0; if mstatus='1. Yes' then mstatusn=1;
      arean=0;    if area ='1. Highly Urban/ urban area' then arean=1;
      sexn=0;     if sex='1. M' then sexn=1;
run;

proc nlmixed data=claims3;
      eta = b_0 + b_car_use*car_usen + b_mstatus*mstatusn +
                  b_area*arean + b_lincome*lincome + b_sex*sexn;
      lambda = exp(eta);
      loglike = - lambda + clm_freq*log(lambda) - log(fact(clm_freq)) ;
      model clm_freq ~ general(loglike);
      * same results if ll is hardcoded;
/*    model clm_freq ~ poisson(lambda); */
run;
```

or;

```
proc nlmixed data=claims2;
      eta = b_0   + b_car_use*(car_use='Commercial')
                  + b_mstatus*(mstatus='1. Yes')
                  + b_area*(area ='1. Highly Urban/ urban area')
                  + b_lincome*lincome
                  + b_sex*(sex='1. M');
      lambda = exp(eta);
      model clm_freq ~ poisson(lambda);
run;
```

The NLMIXED Procedure

Specifications	
Data Set	WORK.CLAIMS3
Dependent Variable	CLM_FREQ
Distribution for Dependent Variable	General
Optimization Technique	Dual Quasi-Newton
Integration Method	None

Dimensions	
Observations Used	2812
Observations Not Used	0
Total Observations	2812
Parameters	12

Fit Statistics	
-2 Log Likelihood	6404.0
AIC (smaller is better)	6428.0
AICC (smaller is better)	6428.1
BIC (smaller is better)	6499.3

Parameter Estimates									
Parameter	*Estimate*	*Standard Error*	*DF*	*t Value*	*Pr > |t|*	*Alpha*	*Lower*	*Upper*	*Gradient*
bll_0	0.4709	0.1302	2812	3.62	0.0003	0.05	0.2156	0.7262	0.267506
bll_1	0.03851	0.05742	2812	0.67	0.5025	0.05	-0.07408	0.1511	0.084439
bll_2	-0.02516	0.05479	2812	-0.46	0.6461	0.05	-0.1326	0.08227	0.034717
bll_3	0.08448	0.1098	2812	0.77	0.4418	0.05	-0.1309	0.2998	0.089421
bll_4	-0.00321	0.008304	2812	-0.39	0.6994	0.05	-0.01949	0.01308	-0.41592
bll_5	0.01517	0.05627	2812	0.27	0.7874	0.05	-0.09517	0.1255	-0.14471
bp_0	1.2705	0.2256	2812	5.63	<.0001	0.05	0.8282	1.7128	-0.20101
bp_1	-0.5539	0.1235	2812	-4.49	<.0001	0.05	-0.7961	-0.3118	0.240928
bp_2	0.1607	0.1125	2812	1.43	0.1532	0.05	-0.05987	0.3813	0.000136
bp_3	-2.0319	0.1554	2812	-13.07	<.0001	0.05	-2.3367	-1.7272	-0.00329
bp_4	0.04199	0.01865	2812	2.25	0.0245	0.05	0.005415	0.07857	-0.41593
bp_5	0.3006	0.1144	2812	2.63	0.0086	0.05	0.07639	0.5248	-0.07819

Contrasts				
Label	*Num DF*	*Den DF*	*F Value*	*Pr > F*
TEST p_0=0.4468	1	2812	0.02	0.8836

Additional Estimates								
Label	*Estimate*	*Standard Error*	*DF*	*t Value*	*Pr > |t|*	*Alpha*	*Lower*	*Upper*
Estimated proportion of 'extra' zeros (theta)	0.4435	0.02280	2812	19.45	<.0001	0.05	0.3988	0.4882
Estimated Conditional Poisson Mean (lambda)	1.6515	0.07387	2812	22.36	<.0001	0.05	1.5067	1.7963
Estimated Unconditional ZIP Mean ((1-p_0)*lambda)	0.9191	0.04280	2812	21.47	<.0001	0.05	0.8352	1.0031
Estimated Unconditional ZIP Variance ((1-p_0)*lambda*(1+p_0*lambda))	1.5923	0.09299	2812	17.12	<.0001	0.05	1.4099	1.7746

Zero-inflated Poisson regression models can also be easily fitted using Proc NLMIXED. The Zero-inflation parameter can be left as a constant, or a second regression equation can be fitted with

the same or different regressors as for the mean parameter, allowing that ZI parameter to vary by group, or even by observation. The mean parameter has a log link, ensuring positivity of the mean, while the ZI parameter has a logit link, ensuring that it remains between zero and one.

Parameter Estimates

	Parameter	Error	DF	t
bll_0	(0.495)	0.119	2,812	(4.170)
bll_1	0.142	0.052	2,812	2.750
bll_2	(0.010)	0.049	2,812	(2.040)
bll_3	1.211	0.097	2,812	12.460
bll_4	0.015	0.007	2,812	(2.100)
bll_5	0.053	0.051	2,812	(1.040)
bp_0	(0.198)	0.058	2,812	(3.390)

Adding regressors as for the ZI parameter.

```
proc nlmixed data=claims3;
  parameters  bll_0=0 bll_1=0 bll_2=0 bll_3=0 bll_4=0 bll_5=0
                      bp_0=0  bp_1=0 bp_2=0 bp_3=0 bp_4=0 bp_5=0;

  eta_prob = bp_0 + bp_1*car_usen + bp_2*mstatusn + bp_3*arean
                  + bp_4*lincome + bp_5*sexn;
  p_0 = exp(eta_prob)/(1 + exp(eta_prob));
 eta_lambda = bll_0 + bll_1*car_usen + bll_2*mstatusn + bll_3*arean
                      + bll_4*lincome + bll_5*sexn;
  lambda = exp(eta_lambda);
 if clm_freq=0 then loglike = log(p_0 + (1-p_0)*exp(-lambda));
               else loglike = log(1-p_0) + clm_freq*log(lambda)
                                  -lambda - lgamma(clm_freq+1);
  model clm_freq ~ general(loglike);
  estimate "Estimated proportion of 'extra' zeros (theta)" p_0;
  estimate 'Estimated Conditional Poisson Mean (lambda)' lambda;
  estimate 'Estimated Unconditional ZIP Mean ((1-p_0)*lambda)'
                                                (1-p_0)*lambda;
  estimate 'Estimated Unconditional ZIP Variance
      ((1-p_0)*lambda*(1+p_0*lambda))' (1-p_0)*lambda*(1+p_0*lambda);

  predict (1-p_0)*lambda out = lambda_hat ;
  title 'ZIP regression model';
run;
```

ZIP regression model

The NLMIXED Procedure
Specifications

Dependent Variable	CLM_FREQ
Distribution for Dependent Variable	General
Optimization Technique	Dual Quasi-Newton

Dimensions
Fit Statistics

-2 Log Likelihood	6404.0

```
AIC (smaller is better)          6428.0
AICC (smaller is better)         6428.1
BIC (smaller is better)          6499.3
```

Parameter Estimates

Parameter	Estimate	Error	DF	t	Pr	Lower	Upper	Gradient
bll_0	0.469	0.130	2812	3.6	0.0003	0.2136	0.7241	0.0525
bll_1	0.037	0.057	2812	0.65	0.5172	-0.0754	0.1498	-0.0033
bll_2	-0.024	0.055	2812	-0.43	0.6661	-0.1311	0.0838	0.1584
bll_3	0.085	0.110	2812	0.78	0.4371	-0.1300	0.3007	0.0884
bll_4	-0.003	0.008	2812	-0.38	0.7062	-0.0194	0.0132	0.5054
bll_5	0.015	0.056	2812	0.27	0.7863	-0.0951	0.1256	-0.0293
bp_0	1.277	0.226	2812	5.66	<.0001	0.8351	1.7197	0.0822
bp_1	-0.559	0.124	2812	-4.52	<.0001	-0.8018	-0.3169	-0.0772
bp_2	0.165	0.113	2812	1.46	0.1434	-0.0560	0.3855	0.4239
bp_3	-2.035	0.156	2812	-13.08	<.0001	-2.3400	-1.7299	0.0725
bp_4	0.042	0.019	2812	2.24	0.0254	0.0051	0.0782	0.1624
bp_5	0.298	0.114	2812	2.61	0.0092	0.0740	0.5225	-0.3611
Prop Extra 0's	0.445	0.023		19.52	0.0001			

Appendix F

Other Count Datasets

Table F.1 Bailey and Simon Data

ID	Class #	Class	Merit	Exposure	Earned	Claims	Frequency/
1	1	NoYoungMale	A	2,757,520	159,108,000	217,151	0.079
2	5	MarriedYoungMale	A	64,130	5,349,000	6,560	0.102
3	1	NoYoungMale	X	130,706	7,910,000	13,792	0.106
4	2	NonPrincipYoungMale	A	130,535	11,840,000	14,506	0.111
5	1	NoYoungMale	Y	163,544	9,862,000	19,346	0.118
6	5	MarriedYoungMale	x	4,039	345,000	487	0.121
7	5	MarriedYoungMale	Y	4,869	413,000	613	0.126
8	3	Business	A	247,424	25,846,000	31,964	0.129
9	1	NoYoungMale	B	273,944	17,226,000	37,730	0.138
10	2	NonPrincipYoungMale	X	7,233	712,000	1,001	0.138
11	4	YoungMale	A	156,871	18,450,000	22,884	0.146
12	2	NonPrincipYoungMale	Y	9,726	944,000	1,430	0.147
13	5	MarriedYoungMale	B	8,601	761,000	1,291	0.150
14	2	NonPrincipYoungMale	B	21,504	1,992,000	3,421	0.159
15	3	Business	X	15,868	1,783,000	2,695	0.170
16	4	YoungMale	y	21,089	2,523,000	3,618	0.172
17	4	YoungMale	X	17,707	2,130,000	3,054	0.172
18	3	Business	Y	20,369	2,281,000	3,546	0.174
19	4	YoungMale	B	56,730	6,608,000	11,345	0.200
20	3	Business	B	37,666	4,129,000	7,565	0.201

Table F.1.a Chi-Square Test Based on Bailey and Simon Data

Poisson/ZIP	7.8E+16
Negative Binomial	6,672,651
ZINB	6,107,153

Table F.2 Wang, Cockburn and Puterman (1998) Patents data

Obs	Company	Patents	RDS	lgRD
1	ABBOTT LABORATORIES	42	0.0549	4.0869
2	AFFILIATED HOSPITAL PRDS	1	0.0032	-2.0794
3	ALBERTO-CULVER CO	3	0.0078	0.1187
4	ALCON LABORATORIES	2	0.0803	1.8796
5	ALLERGAN PHARMACEUTICALS INC	3	0.0686	1.1033
6	ALZA CORP-CL A	40	3.3319	2.0794
7	AMERICAN HOME PRODUCTS CORP	60	0.0243	4.0953
8	AMERICAN HOSPITAL SUPPLY	30	0.0128	2.8333
9	AMERICAN STERILIZER CO	7	0.0252	1.3915
10	AVON PRODUCTS	3	0.0094	2.6048
11	BARD(C.R.) INC	5	0.0146	0.7957
12	BAXTER TRAVENOL LABORATORIES	59	0.0496	3.5207
13	BECTON, DICKINSON & CO	26	0.0395	3.0001
14	BENTLEY LABORATORIES	3	0.0780	0.5371
15	BOCK DRUG-CL A	0	0.0171	0.7761
16	BRISTOL-MYERS CO	66	0.0347	4.2338
17	CARTER-WALLACE INC	0	0.0569	2.2178
18	CAVITRON CORP	8	0.1095	0.8510
19	CHATTEM INC	2	0.0190	-0.1567
20	CHESEBROUGH-POND'S INC	4	0.0084	1.8358
21	CLINICAL SCIENCES INC	0	0.1003	-1.6045
22	CODE LABORATORIES INC	0	0.0623	0.7071
23	CONCEPT INC	3	0.0707	-0.9916
24	COOPER LABORATORIES	6	0.0359	1.2296
25	DATASCOPE CORP	3	0.0596	-0.5310
26	DEL LABORATORIES INC	0	0.0076	-1.2310
27	DENTSPLY INTERNATIONAL INC	6	0.0185	0.9270
28	DESERET PHARMACEUTICAL	2	0.0080	-1.1332
29	DYNATECH CORP	3	0.0640	-0.0419
30	ELECTRO CATHETER CORP	0	0.0780	-1.8326
31	EVEREST & JENNINGS INTL	1	0.0025	-1.8264

Obs	Company	Patents	RDS	lgRD
32	FABERGE INC	1	0.0040	-0.1985
33	FOREST LABORATORIES INC	0	0.0329	-1.7838
34	GILLETTE CO	25	0.0234	3.5525
35	GUARDIAN CHEMICAL CORP	2	0.0387	-2.5639
36	HELENE CURTIS INDUSTRIES	4	0.0133	0.4523
37	ICN PHARMACEUTICALS INC	1	0.0324	1.0529
38	INSTRUMENTATION LABS INC	1	0.0882	1.4873
39	INTL FLAVORS & FRAGRANCES	51	0.0587	2.7793
40	JOHNSON & JOHNSON	105	0.0446	4.7233
41	JOHNSON PRODUCTS	1	0.0131	-0.6444
42	KEY PHARAMACEUTICALS INC	0	0.0160	-2.9565
43	LA MAUR INC	0	0.0143	-0.9545
44	LILLY (ELI) & CO	166	0.0843	4.7278
45	MALLINCKRODT INC	8	0.0320	2.1831
46	MARION LABORATORIES	6	0.0599	1.5773
47	MERCK & CO	173	0.0821	4.9152
48	????	25	0.0535	3.1807
49	MINE SAFETY APPLIANCES CO	14	0.0226	1.4036
50	NARCO SCIENTIFIC INC	3	0.0397	1.0043
51	NESTLE-LEMUR CO	0	0.0103	-2.3330
52	NEWPORT PHARMACEUTICALS INTL	0	0.7159	-0.1815
53	NOXELL CORP	2	0.0107	0.2670
54	PFIZER INC	93	0.0467	4.4785
55	PURITAN-BENNETT CORP	3	0.0369	0.7105
56	REDKEN LABORATORIES	2	0.0316	0.2979
57	RESEARCH INDUSTRIES CORP	0	0.0355	-2.8647
58	REVLON INC	5	0.0166	2.7622
59	RICHARDSON-MERRELL INC	23	0.0417	3.4383
60	ROBINS (A.H.) CO	11	0.0447	2.5439
61	RORER GROUP	13	0.0401	2.4436
62	SCHERER (R.P.)	0	0.0050	-0.4125
63	SCHERING-PLOUGH	90	0.0618	3.9865

Obs	Company	Patents	RDS	lgRD
64	SEARLE (G.D.) & CO	63	0.0690	3.9620
65	SMITHKLINE CORP	112	0.0813	4.0029
66	SQUIBB CORP	115	0.0409	3.9051
67	STERLING DRUG INC	48	0.0331	3.5909
68	SYBRON CORP	15	0.0323	2.9242
69	SYNTEX CORP	69	0.0859	3.1707
70	TECHNICARE CORP	4	0.0591	1.8089

Table F.2.a Chi-Square Test Based on Patent Data

Distribution	Chi Square
Poisson	7.10E+84
Negative Binomial	584,069
ZIP	6.70E+71
ZINB	249

Table F.3 Deb and Trivedi Hospital Stay Counts

dsn=dt, yvar=hosp, dist=poisson

poisson distribution with mean(xbar)=sample mean= 0.29596

y	Frequency	Percent	Test Percent	Cumulative Frequency	Cumulative Percent
0	3541	80.370	74.380	3541	80.37
1	599	13.600	22.010	4140	93.96
2	176	3.990	3.260	4316	97.96
3	48	1.090	0.320	4364	99.05
4	20	0.450	0.020	4384	99.5
5	12	0.270	0.000	4396	99.77
6	5	0.110	0.000	4401	99.89
7	1	0.020	0.000	4402	99.91
8	4	0.090	0.000	4406	100

Table F.3.a Chi-Square Test Based on Hospital Visit Data

Distribution	Chi Square
Poisson	3.40E+06
Negative Binomial	19,302
ZIP	2.,925
ZINB	25

Table F.4 Deb and Trivedi Office Visit Data

y	Frequency	Percent	Test Percent	Cumulative Frequency	Cumulative Percent
0	683	15.5	0.31	683	15.5
1	481	10.92	1.79	1164	26.42
2	428	9.71	5.18	1592	36.13
3	420	9.53	9.97	2012	45.67
4	383	8.69	14.39	2395	54.36
5	338	7.67	16.62	2733	62.03
6	268	6.08	15.99	3001	68.11
7	217	4.93	13.19	3218	73.04
8	188	4.27	9.52	3406	77.3
9	171	3.88	6.11	3577	81.18
10	128	2.91	3.53	3705	84.09
11	115	2.61	1.85	3820	86.7
12	86	1.95	0.89	3906	88.65
13	73	1.66	0.4	3979	90.31
14	76	1.72	0.16	4055	92.03
15	53	1.2	0.06	4108	93.24
16	47	1.07	0.02	4155	94.3
17	48	1.09	0.01	4203	95.39
18	30	0.68	0	4233	96.07
19	24	0.54	0	4257	96.62
20	16	0.36	0	4273	96.98
21	18	0.41	0	4291	97.39
22	16	0.36	0	4307	97.75
23	10	0.23	0	4317	97.98
24	12	0.27	0	4329	98.25
25	3	0.07	0	4332	98.32
26	9	0.2	0	4341	98.52
27	7	0.16	0	4348	98.68
28	4	0.09	0	4352	98.77
29	3	0.07	0	4355	98.84
30	4	0.09	0	4359	98.93
>30	47	1.04	0	4406	100

Table F.4.a Chi-Square Test Based on Office Visit Data

Distribution	Chi Square
Poisson	2.02E+67
Negative Binomal	2,856
ZIP	2.33E+01
ZINB	4,224

Table F.5 Ridout and Demetrio Apple Shoot Counts

Frequency distributions of the number of roots									
by 270 shoots of the apple cultivar Trajan									
BAP (muM)	**Photoperiod**								**All**
	8				**16**				
	2	**4**	**9**	**18**	**2**	**4**	**9**	**18**	
No. of roots	0	0	0	2	15	16	12	19	64
0									
1	3	0	0	0	0	2	3	2	10
2	2	3	1	0	2	1	2	2	13
3	3	0	2	2	2	1	1	4	15
4	6	1	4	2	1	2	2	3	21
5	3	0	4	5	2	1	2	1	18
6	2	3	4	5	1	2	3	4	24
7	2	7	4	4	0	0	1	3	21
8	3	3	7	8	1	1	0	0	23
9	1	5	5	3	3	0	2	2	21
10	2	3	4	4	1	3	0	0	17
11	1	4	1	4	1	0	1	0	12
12	0	0	2	0	1	1	1	0	5
13	1	1	.	.	.	.	.	.	2
14	.	.	2	1	.	.	.	.	3
17	1	.	.	.	.	.	.	.	1
All	30	30	40	40	30	30	30	40	270

Table F.5.a Chi-Square Test Based on Ridout and Demetrio Apple Shoot Data

Distribution	Chi Square
Poisson	2,694
NB	131
ZIP	76
ZINB	19

Table F.6 Long Biochemists Data

The FREQ Procedure

y	Frequency	Percent	Test Percent	Cumulative Frequency	Cumulative Percent
0	275	30.05	19.28	275	30.05
1	246	26.89	31.74	521	56.94
2	178	19.45	26.12	699	76.39
3	84	9.18	14.33	783	85.57
4	67	7.32	5.90	850	92.90
5	27	2.95	1.94	877	95.85
6	17	1.86	0.53	894	97.70
7	21	2.30	0.13	915	100.00

Table F.6.a Chi-Square Test Based on Biochemists Data

Chi-Square Test for Specified Proportions	
Chi-Square	476.5913
DF	7
Pr > ChiSq	<.0001
WARNING: 25% of the cells have expected counts less than 5. Chi-Square may not be a valid test.	
Sample Size = 915	

Appendix G

A Simple Procedure for Fitting the ZIP Model

A procedure to solve for the ZIP parameters can be set up in Microsoft Excel. The illustration used in this paper optimizes the minimum distance procedure as set forth in Hogg and Klugman (1984), but other statistics can be optimized.

$$\text{(G.1)} \qquad \min\left(\sum_k w(k)[F_n(k) - F(k)]^2\right)$$

$w(k)$ = weight, $F_n(k)$ = Empirical DF, $F(k)$ = Fitted DF

Hogg and Klugman suggest using a weight of: $w(k) = \dfrac{n_k}{F_n(k)(1-F_n(k))}$

Table 2.2 displays the spreadsheet setup for the parameter estimation. The parameters phi and lambda have been initialized to those of the Poisson distribution, i.e., no structural zeros, so phi is zero. The sum of column (7) is to be minimized.

Table G.1
Calculation of Zero-Inflated Poisson Parameters: Initialization

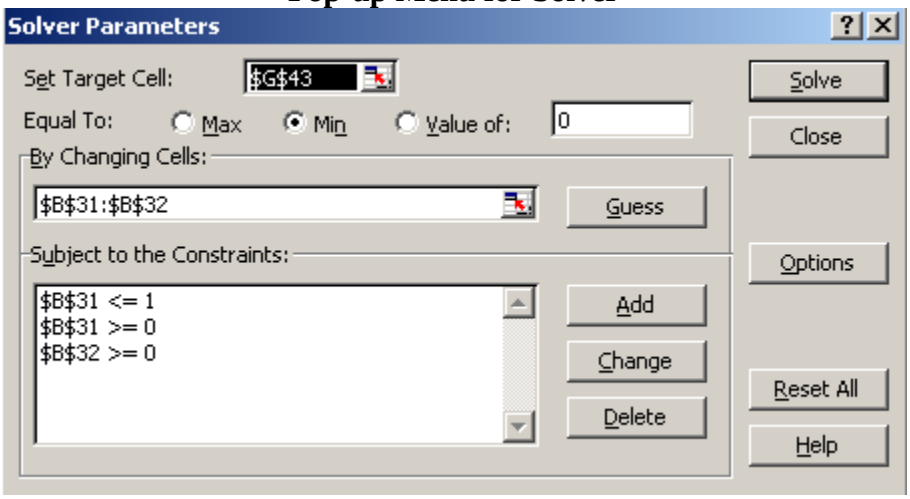
Phi	0					
Lambda	0.82					
No Claims	Actual Count	P(X=x) (2)/SUM(2)	Theoretical P(X=x)	Weight (1)/((2)(1-(2))	Squared Deviation ((3)-(4))^2	Wt Squared Deviation (5)*(6)
(1)	(2)	(3)	(4)	(5)	(6)	(7)
0	1,706	0.607	0.44043	7,149	0.02764	197.6
1	351	0.125	0.36115	3,213	0.05585	179.5
2	408	0.145	0.14807	3,289	0.00001	0.0
3	268	0.095	0.04047	3,108	0.00301	9.3
4	74	0.026	0.00830	2,888	0.00032	0.9
5	5	0.002	0.00136	2,817	0.00000	0.0
Sum	2,812				0.08683	387.4

Excel provides the solver function to solve[12] nonlinear optimization problems such as this one. Solver uses a numerical algorithm, such as gradient descent, to solve nonlinear problems. Figure 2.2 displays the pop-up menu that is used with Solver. The menu requires the user to identify a target cell to optimize (here the sum of the weighted squared deviations), the input cells containing the parameters to be estimated and whether the optimization is a minimization or maximization.

[12] Please note you must load the solver add-in to use solver. This can be done from the tools menu, but requires the Microsoft Office disk.

Optionally, the user can specify constraints on the parameters (i.e., for instance phi must be greater than or equal to zero and less than or equal to 1.

Figure G.1
Pop-up Menu for Solver



The Poisson parameters fit with Excel solver are displayed in Table 2.3. The table indicates that on average, 54% of the records have structural zeros. For the remaining policyholders, the mean claim frequency over a five-year period is approximately 1.9. Figure 2.3 compares the negative binomial to the zero-inflated Poisson. The ZIP model appears to provide a better fit to the data.

Table G.2
Fitted Zero-Inflated Poisson

Phi	0.5359					
Lambda	1.9194					
No Claims	Actual Count	P(X=x) (2)/SUM(2)	Theoretical P(X=x)	Weight (1)/((3)(1-(3))	Squared Deviation ((3)-(4))^2	Wt Squared Deviation (5)*(6)
(1)	(2)	(3)	(4)	(5)	(6)	(7)
0	1706	0.607	0.60402	7149	0.00001	0.1
1	351	0.125	0.13066	3213	0.00003	0.1
2	408	0.145	0.12540	3289	0.00039	1.3
3	268	0.095	0.08023	3108	0.00023	0.7
4	74	0.026	0.03850	2888	0.00015	0.4
5	5	0.002	0.01478	2817	0.00017	0.5
Sum	2,812				0.00097	3.0

Table G.3
Fitted Zero-Inflated Negative Binomial Model

phi	0					
r	1					
p	0.4561					
No Claims	Actual Count	P(X=x) (2)/SUM(2)	Theoretical P(X=x)	Weight (1)/((3)(1-(3))	Squared Deviation ((3)-(4))^2	Wt Squared Deviation (5)*(6)
(1)	(2)	(3)	(4)	(5)	(3290)	(3291)
0	1706	0.607	0.4561	7149	0.00048	1.5
1	351	0.125	0.2481	3213	0.00018	0.5
2	408	0.145	0.1349	3289	0.00040	1.1
3	268	0.095	0.0734	3108	0.00000	0.0
4	74	0.026	0.0399	2888	0.00000	0.0
5	5	0.002	0.0217	2817	0.00000	0.0
Sum					0.00106	3.1

Table G.4
Fitted Hurdle Poisson Model

phi	0.599					
lambda	1.9286					
No Claims	Actual Count	P(X=x) (2)/SUM(2)	Theoretical P(X=x)	Weight (1)/((3)(1-(3))	Squared Deviation ((3)-(4))^2	Wt Squared Deviation (5)*(6)
(1)	(2)	(3)	(4)	(5)	(6)	(7)
0	1706	0.607	0.5990	7149	0.00006	0.4
1	351	0.125	0.1124	3213	0.00015	0.5
2	408	0.145	0.1084	3289	0.00135	4.4
3	268	0.095	0.0697	3108	0.00066	2.0
4	74	0.026	0.0336	2888	0.00005	0.2
5	5	0.002	0.0130	2817	0.00013	0.4
Sum	2,812				0.00239	7.9

5. REFERENCES

[1.] Anderson, Duncan, and Sholom Feldblum, Claudine Modlin, Doris Schirmacher, Ernesto Schirmacher, Neeza Thandi. A Practioner's Guide to GLMs, CAS Exam Study Note, **2005**.

[2.] Arnold, B., *Pareto Distributions*, International Co-Operative Publishing House, **1983.**

[3.] Bailey, R. and Simon, L., Two Studies in Automobile Insurance Ratemaking, *PCAS* XLVII, **1960,** pp. 1-19.

[4.] Bailey, R. and L. Simon, An Actuarial Note on the Credibility of Experience of a Single Private Passenger Car, *PCAS* **1959**, pp. 159-164.

[5.] Brown, Robert L., "Minimum Bias With Generalized Linear Models," *Proceedings of the Casualty Actuarial Society (PCAS)*, **1988**, 45, p187-217.

[6.] Clark, D. and C. Thayer, "A Primer on the Exponential Family of Distributions," *CAS Discussion Paper Program*, **2004.**

[7.] Czado C., and A. Min, "Zero-inflated Generalized Poisson Regression Models: Asymptotic Theory and Applications," Center for Mathematical Sciences, Munich University of Technology.

[8.] Deb, Partha and Pravin K. Trivedi, "Demand for Medical Care by the Elderly: A Finite Mixture Approach," *Journal of Applied Econometrics*, Vol. 12, No. 3, **1997**, pp. 313-336.

[9.] de Jong, P. and Heller, G., *Generalized Linear Models for Insurance Data*, Cambridge, **2008.**

[10.] De Ville, *Decision Trees for Business Intelligence and Data Mining*, SAS Press, **2006.**

[11.] Fish, J., C. Gallagher, and H. Monroe, "An Iterative Approach to Classification Analysis," *Casualty Actuarial Society Discussion Paper Program*, pp. 237 – 281, **1990.**

[12.] Faraway, J., *Extending the Linear Model with R*, Chapman and Hall, **2006.**

[13.] Grogger, Jeffery and Richard T. Carson, "Models for Truncated Counts," *Journal of Applied Econometrics*, 1991, v6, p225-238.

[14.] He, B., M. Xie, T. N. Goh, and K. L. Tsui, "Control Charts Based on Generalized Poisson Model for Count Data." Technical Report, Logistics Institute of Georgia Tech., **2002**, http://www.isye.gatech.edu/research/files/tsui-2002-04.pdf.

[15.] Heilbron, David C., "Zero-Altered And Other Regression Models For Count Data With Extra Zeros," *Boimetrical Journal*, **1994**, vol. 36, pp. 531-547.

[16.] Hilbe, Joesph M., *Negative Binomial Regression,* Cambridge University Press, **2007**.

[17.] Hogg, R. and S. Klugman, *Loss Distributions*, Wiley, **1984.**

[18.] Ismail, Noriszura and Abdul Aziz Jemain, "Handling Overdispersion with Negative Binomial and Generalized Poisson Regression Models," CAS Winter *Forum* including the Ratemaking Call Papers, **2007**.

[19.] Jackman, Simon with contributions from Alex Tahk, Achim Zeileis, Christina Maimone, and Jim Fearon, pscl: Political Science Computational Laboratory, Stanford University, **2007.**

[20.] Johnson, N., S. Kotz, and A. Kemp, *Univariate Discrete Distributions*, 2nd Ed., Wiley, **1992.**

[21.] Lambert, D, "Zero-Inflated Poisson With An Application To Manufacturing Defects," *Techonometrics*, February **1992**, pp 1-14.

[22.] Long, J. Scott. Regression *Models for Categorical and Limited Dependent Variables*, (Thousand Oaks, California: Sage, **1997**).

[23.] McCulloch, Peter and John A. Nelder, *Generalized Linear Models*, Chapman and Hall, **1989**.

[24.] Mullahy, John, "Specification And Testing Of Some Modified Count Data Models," *Journal of Econometrics*, **1986**, vol. 33, pp 341-365.

[25.] Meyers, Glenn, "Estimating Loss Costs at the Address Level," PowerPoint presentation at the CAS Predictive Modeling Seminar, **2007**, https://www.casact.org/education/specsem/f2007/handouts/meyers.ppt#295,1.

[26.] Press W., B. Flannery, S. Tuekolsky, and W. Vetterling, *Numerical Recipes*, Cambridge University Press, **1989**.

[27.] Refaat, M., *Data Preparation With Data Mining,* SAS Institute, **2007.**

[28.] Ridout, M., C. Demetrio, and J. Hinde, "Models for Count Data with Many Zeros," presentation a International Biometric Conference, **1998**.

[29.] Shaw, Daigee, "On-Site Samples' Regression, Problems Of Non-Negative Integers, Truncation And Endogenous Stratification," *Journal of Econometrics*, **1988**, vol. 37, no. 2, pp 2110-233.

[30.] Simon, L., "An Introduction to the Negative Binomial Distribution and its Application," *PCAS* **1962**, pp. 1-8.
[31.] SPSS Technical Support, "CHAID and Exhaustive CHAID Algorithms," www.spss.com.
[32.] Unknown, CHAID like Discrimination Classification, http://citeseer.ist.psu.edu/151555.html.
[33.] Van den Broek, Jan, "A Score Test For Zero Inflation In A Poisson Distribution," *Biometrics*, **1995**, vol. 51, no. 2, pp 738-743.
[34.] Venables, W. N. and B. Ripley, *Modern Applied Statistics with S,* 4th ed., Springer, **2002**.
[35.] Wang, P., I.M. Cockburn, and M. L. Puterman, "Analysis of Patent Data— A Mixed-Poisson-Regression-Model Approach," *Journal of Business & Economic Statistics*, **1998**, 16, 1, p27-41, https://www.amstat.org/publications/jbes/index.cfm?fuseaction=wang1998.
[36.] Winklemann, Rainer and Klauss F. Zimmermann, "Recent Developments In Count Data Modelling: Theory And Application," *Journal of Economic Surveys*, **1995**, vol. 9, pp 1-24.
[37.] Yip, Karen C. H. and Kelvin K. W. Yau, "On Modeling Claim Frequency Data In General Insurance With Extra Zeros," *Insurance: Mathematics and Economics*, **2005**, vol. 36, no. 2, pp 153-163.

Acknowledgment

The authors gratefully acknowledge the editorial assistance of Jane Taylor. They also thank their reviewers for their many helpful comments.

Abbreviations and notations

Collect here in alphabetical order all abbreviations and notations used in the paper

ZIP: Zero-Inflated Poisson
ZINB Zero-Inflated negative binomial
CHAID Chi-Squared Automatic Interaction Detection
GLM, generalized linear models
CAS Casualty Actuarial Society

Biographies of the Authors

Matthew Flynn, Ph.D. is Director of Research at ISO Innovative Analytics.

Louise Francis is a Consulting Principal at Francis Analytics and Actuarial Data Mining, Inc. She is involved in data mining projects as well as conventional actuarial analyses. She has a BA degree from William Smith College and an MS in Health Sciences from SUNY at Stony Brook. She is a Fellow of the CAS and a Member of the American Academy of Actuaries. She is the 2008-09 through 2010-11 Vice President of R&D for the CAS. She is a four-time winner of the Data Quality, Management and Technology call paper prize.