Rate Making using Frequency Severity Modeling in General InsuranceProduct with Excess Zero Count

Contents

Introduction

Accidents occur every day. Some more severe than others. We cannot prevent accidents from happening. We can, however, protect ourselves against great financial loss if one does occur. Insurance is a way of protection against great economic losses. As a customer one is looking for an insurance policy that covers as much as possible at a low cost. As an insurer you want to make money, and at the same time keep and gain customers. If the cost of insurance is too high, customers will go to a company that is cheaper. In that way you are losing customers to the competition. If the price is too low, the insurer can risk more money going out, than coming in. As a result, the company loses money or, in the worst-case scenario, becomes insolvent. Individuals that are involved in many accidents, often get high premiums. When the price of insurance is too low, you will also risk attracting customers that have a high rate of claims, since the price they would get with other companies would be higher. This would fuel the initial problem, by increasing the number of claims. Therefore, pricing of insurance is crucial.

Insurance is an arrangement designed to protect a policyholder from financial losses and can roughly be divided into two main types. The first one is called life insurance and is related to the risk of an individual life, for instance, death, disability and retirement. The second is non-life insurance and deals with property losses or damages. Examples are insurance for home, travel and automobile. Pricing methods of the two types of insurance differ from each other, and our focus will be on the latter case. More specifically, we will concentrate on automobile insurance.

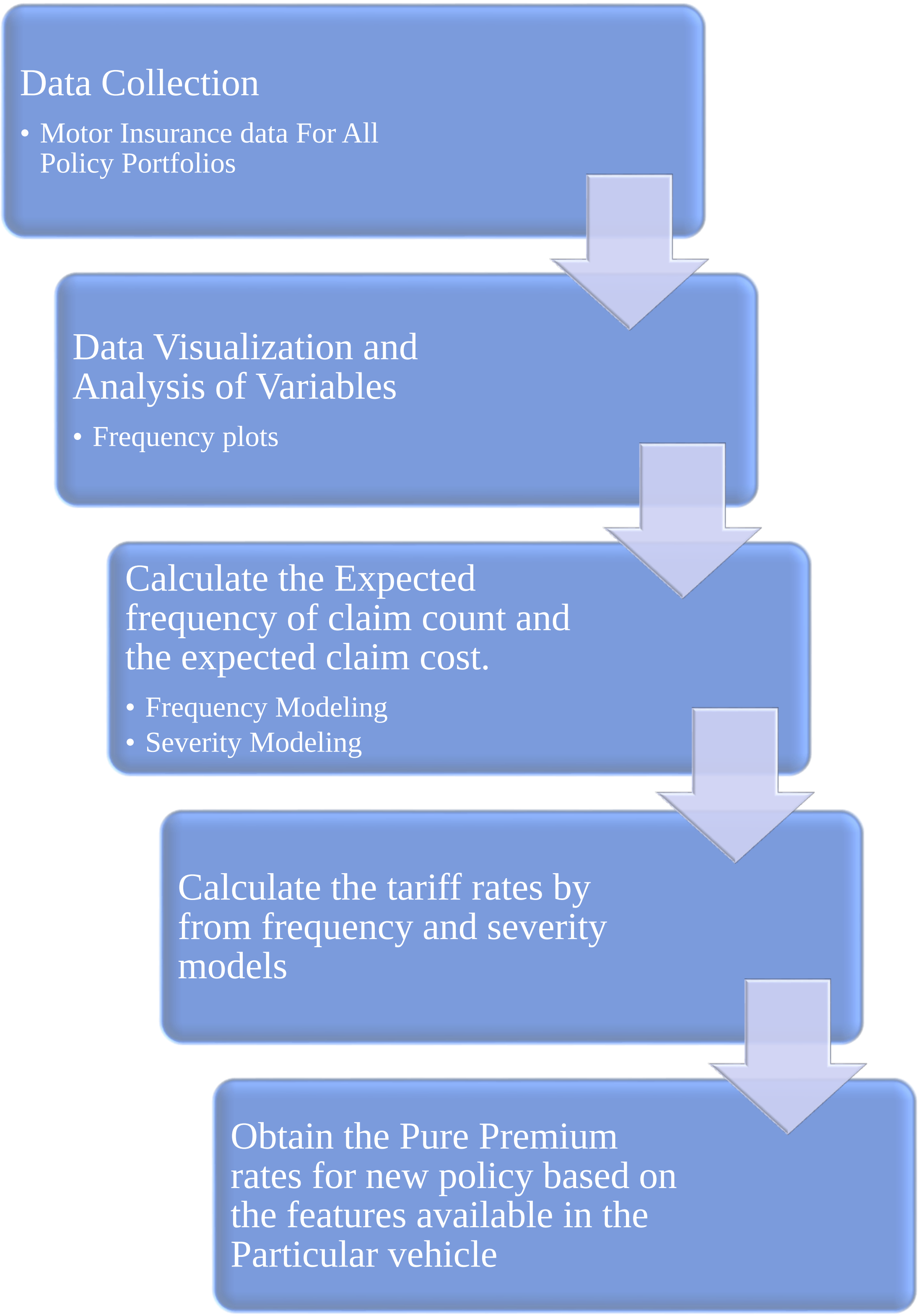
An insurance policy is the financial contract between an insurer (e.g. an insurance company) and a policyholder. The insurer takes all or part of the risk and demands an agreed amount of money, called insurance premium. This could be either a series of payments over time or a single payment. For the insurance premium, overhead costs (administrative expenses, capital costs etc.) and profits are taken into account. The part of the insurance premium that corresponding to the risk is called pure premium. It represents the expected pay-out for reported claims that occurred during the policy period.

Motivation

Being a part of Actuarial studies pricing has been always an important part of the product accessing. In General Insurance the pricing techniques are gaining more importance because of the various factors involved in different claims and the policy holders.

Statistical Methodologies are getting Introduced along with classical pricing techniques. Here the distributions used are not a new concept. Our Aim in this project is to point out the several distributions and also some modified ones to gain perfect accuracy. Our work mainly focus on a very common problem in General Insurance Product Pricing issue, i.e. the excess zero count.

Outline of Study



Chapter 2

Frequency Modeling

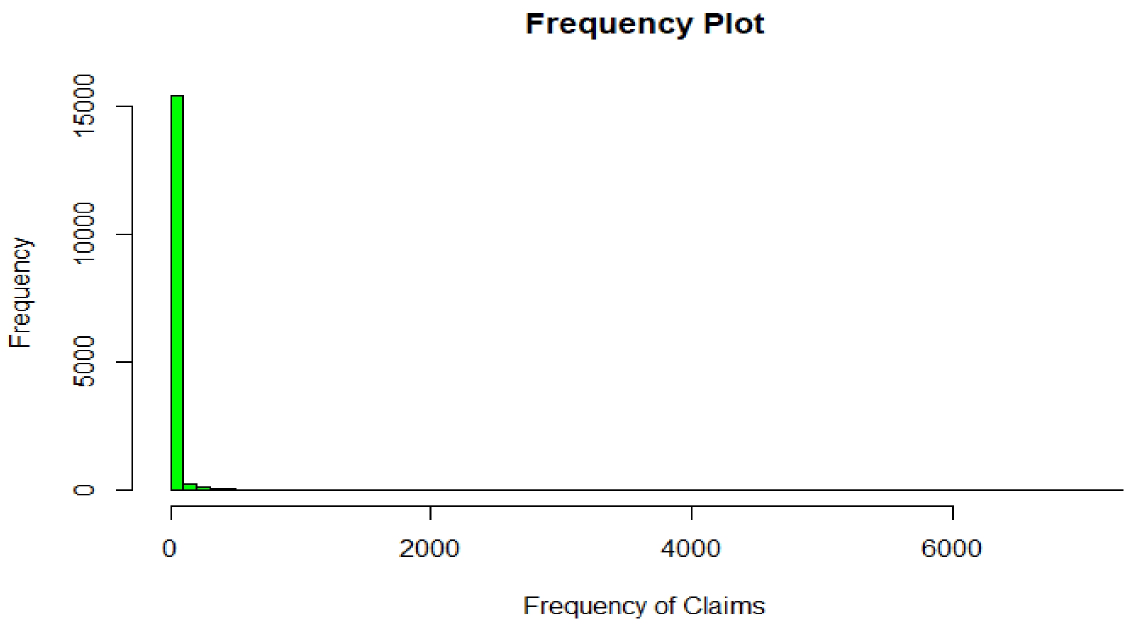
Intuition

Claim frequency and claim severity are the main two risk drivers in general insurance. Many insurance models make restrictive distribution assumptions on latter random variables and even assume them to be independent. The premium charged to a customer is given by the multiplication of the expected claim frequency and severity.

Data Description

Here We have considered the Motor Insurance data of United India Insurance Company. As far as the data is concerned, this is a product basis data, not a customer basis data. Data contains the product details of Motor Insurance products till 2016. The Exposure period is not known.

Data Visualization



Frequency Modeling

Due to this current trend in insurance, generalized linear models (GLMs) have become a popular statistical tool to analyse and model claim frequency and severity.

Generalised Linear Model

Generalized linear modelling is used to assess and quantify the relationship between a response variable and explanatory variables. The modelling differs from ordinary regression modelling in two important respects:

i. The distribution of the response is chosen from the exponential family. Thus, the distribution of the response need not be normal or close to normal and may be explicitly non-normal.
ii. A transformation of the mean of the response is linearly related to the explanatory variables.

A consequence of allowing the response to be a member of the exponential family is that the response can be, and usually is, heteroskedastic. Thus, the variance will vary with the mean which may in turn vary with explanatory variables. This contrasts with the homoscedastic assumption of normal regression.

Generalized linear models are important in the analysis of insurance data. With insurance data, the assumptions of the normal model are frequently not applicable. For example, claim sizes, claim frequencies and the occurrence of a claim on a single policy are all outcomes which are not normal. Also, the relationship between outcomes and drivers of risk is often multiplicative rather additive.

Exponential dispersion family

A probability distribution is a member of the exponential dispersion family if the density function can be expressed by

$$\boldsymbol{f(y_i;\theta_i,\varphi) = \mathbf{exp}\left(\frac{y_{i\theta_i - b(\theta_i)}}{a(\varphi)} + c(y_i,\varphi)\right)}$$

Components of Generalised Linear Model

random component: The conditional distribution of the ${}^{y_i}/_{x_i}$, with mean $\varepsilon(y_i) = \mu_i$. Under classical assumptions, this is independent, normal with constant variance σ^2, i.e., $y_{i\widetilde{iid}}N(\mu_i,\sigma^2)$. In the GLM, the probability distribution of the y_i can be any member of the exponential family, including the normal, Poisson, binomial, gamma, and others. Subsequent work has extended this framework to include multinomial distributions and some non-exponential families such as the negative binomial distribution.

systematic component: The idea that the predicted value of y_i itself is a linear combination of the regressors is replaced by that of a *linear predictor* η, that captures this aspect of linear models,

link function: The connection between the mean of the response, μ_i, and the linear predictor, η_i is specified by the *link function*, $g(\bullet)$, giving

$$g(\mu_i) = \eta_i = x_i^T \beta.$$

Common Link function:

Link Name	Function: $\eta_i = g(\mu_i)$	Inverse: $\mu_i = g^{-1}(\eta_i)$
Identity	μ_i	η_i
square-root	$\sqrt{\mu_i}$	η_i^2
Log	$log_e(\mu_i)$	$\exp(\eta_i)$
Inverse	μ_i^{-1}	η_i^{-1}
inverse-square	μ_i^{-2}	$\eta_i^{-1/2}$
Logit	$log_e \frac{\mu_i}{1-\mu_i}$	$\frac{1}{1+\exp(-\eta_i)}$
Probit	$\varphi^{-1}(\mu_i)$	$\varphi(\eta_i)$
log-log	$-log_e[-log_e(\mu_i)]$	$\exp[-\exp(-\eta_i)]$
comp.log-log	$log_e[-log_e(1-\mu_i)]$	$1 - exp[-\exp(\eta_i)]$

Common Distributions used for GLM:

Family	Notation	Canonical link	Range of y	Variance Function, $\mathcal{V}(\mu \mid \eta)$
Gaussian	$N(\mu, \sigma^2)$	identity: μ	$(-\infty, +\infty)$	φ
Poisson	$Pois(\mu)$	$log_e(\mu)$	$0,1, \dots, \infty$	μ
Negative-Binomial	$NBin(\mu, \theta)$	$log_e(\mu)$	$0,1, \dots, \infty$	$\mu + \mu^2/\theta$
Binomial	$Bin(n, \mu)/n$	$logit(\mu)$	$\{0,1, \dots, n\}/n$	$\mu(1-\mu)/n$
Gamma	$G(\mu, \nu)$	μ^{-1}	$(0, +\infty)$	$\varphi\mu^2$
Inverse-Gaussian	$IG(\mu, \nu)$	μ^2	$(0, +\infty)$	$\varphi\mu^3$

Iterative algorithms

One commonly used iterative algorithms for GLMs is the Fisher scoring algorithm. The idea of the algorithm in light of GLMs is based on the second order Taylor expansion of the log-likelihood.

Goodness-of-fit tests

The tests assess the overall performance of a model in reproducing the data. The commonly used measures include the Pearson chi-square and likelihood ratio deviance statistics, which can be seen as weighted sums of residuals.

The *residual deviance* statistic, as in logistic regression and loglinear models, is defined as twice the difference between the maximum possible log-likelihood for the *saturated model* that fits perfectly and maximized log-likelihood for the fitted model. The deviance can be defined as

$$\boldsymbol{D}(\boldsymbol{y}, \widehat{\boldsymbol{\mu}}) = \mathbf{2}[\boldsymbol{log_e}\boldsymbol{\mathcal{L}}(\boldsymbol{y}; \boldsymbol{y}) - \boldsymbol{\mathcal{L}}(\boldsymbol{y}; \widehat{\boldsymbol{\mu}})]$$

Comparing non-nested models

The flexibility of the GLM and its extensions allows us to fit models to the same data using different families and different link functions, and to fit models that allow for overdispersion or that make special provisions for zero counts.

One price paid for this additional versatility is that standard LR tests and F tests (such as provided by anova() and linearHypothesis() in the car package) do not apply to models that are not nested; that is, where one model cannot be represented as a restricted, special case of another.
For models estimated by maximum likelihood, one general route to comparing non-nested models is through the AIC information criterion proposed initially by Akaike (1973) and the related BIC criterion (Schwartz, 1978), based on the fitted log-likelihood function:

$AIC = -2log_e\mathcal{L}+2\text{k}$ $BIC = -2log_e\mathcal{L} + log_e(n)k$	

These both penalize models with larger k, the number of parameters in the model, with BIC adding a greater penalty with larger sample size.

AIC and BIC do not give significance tests for assessing whether one model is significantly “better” than another.

Voung Test

It is based on comparing the predicted probabilities or the pointwise log-likelihoods of the two models and test the null hypothesis that each is equally close to the saturated model, against the alternative that one model is closer.

GLM Models for Count Data

Poisson Model

A fundamental distribution for modelling count data is the Poisson. The prototypical GLM for count data, where the response y_i takes on non-negative values 0, 1, 2, . . ., uses the Poisson family with the log link.

The pmf of Poisson Distribution is $f(y;\lambda) = \frac{\lambda^y}{y!}e^{-\lambda}, \qquad y = 0,1,\ldots$

With the mean and variance $E[Y] = Var[Y] = \lambda$**.** In short, we use the expression $\sim poisson(\lambda)$.

The claim frequency is obtained by dividing the claim counts by the exposure. In our case, as it is a portfolio basis data, the exposure is not available. So we use the Policy Count. It is often more appropriate to use claim frequency for modelling, since the number of policies under different product may be different.

With a log-link, the Poisson GLM becomes

$log\frac{y_i}{t_i} = x_i\beta \;\; \Leftrightarrow \lambda_i = t_i\, exp(x_i\beta), \qquad y_i \sim poisson(\lambda_i)$

Where $\lambda_i = E\left[{}^{y_i}/_{x_i}\right]$, and the term t_i is known as an offset.

Mean Variance Relation

mean	var	ratio
14.57892	13501.94843	926.12791

Models for Over-Dispersed Count Data

In practice, the Poisson model is often very useful for describing the relationship between the mean μi and the linear predictors, but typically underestimates the variance in the data.

The Poisson model requires the mean to be equal to the variance, which is not satisfied for many datasets of interest. When the variance is greater than the mean, the data are said to be over dispersed. In the opposite case, they are said to be under dispersed.

Negative Binomial

The negative binomial distribution has two parameters, and hence, it is more flexible for fitting data compared to the Poisson.

The pmf of Negative Binomial distribution is

$$f(y;\mu,\tau) = \frac{\Gamma\left(y + {}^{1}/_{\tau}\right)}{\Gamma\left({}^{1}/_{\tau}\right)\Gamma(y+1)}\left(\frac{\mu}{\mu+\frac{1}{\tau}}\right)^{y}\left(\frac{{}^{1}/_{\tau}}{\mu+\frac{1}{\tau}}\right)^{{}^{1}/_{\tau}}$$

Where $\tau = {}^{1}/_{k}$ is called dispersion parameter. The mean and variance of y are given by $E[Y] = \mu$ and $var[Y] = \mu + \tau\mu^2$.

The link function used here is same as Poisson with same justification.

Some Specially designed Distributions for Count Data

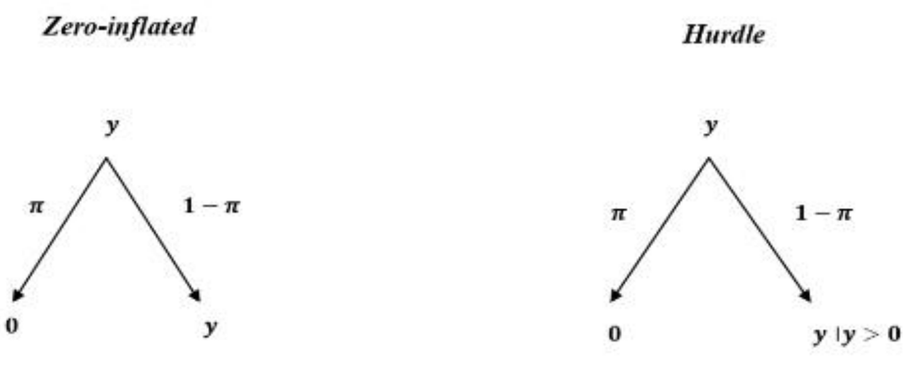
Models for Excess Zero Count

In addition to over dispersion, many sets of empirical data exhibit a greater prevalence of zero counts than can be accommodated by the Poisson or negative-binomial models.

Studies of the distribution of insurance claims often shows large numbers who make no claims because of under-reporting of small claims, policy deductible provisions, and desire to avoid premium increases. Beyond simply identifying this as a problem of lack-of-fit, understanding the reasons for excess zero counts can make a contribution to a more complete explanation of the phenomenon of interest and this requires both new statistical models and visualization techniques.

A statistical formulation of this idea leads to the class of "***zero-inflated***" models described below.

A different form of explanation is that there may be some special circumstance or ***"hurdle"*** required to achieve a positive count, like publishing the master's thesis (such as being driven internally by a personality trait or externally by pressure from a mentor). This idea leads to the class of *hurdle* models that entertain and fit (simultaneously) two separate models: one for the occurrence of the zero counts and one for the positive counts. These two approaches are illustrated as follows.



Zero Inflated Models

Zero-inflated models, introduced by Lambert (1992) as the *zero-inflated Poisson* (ZIP) model, provide an attractive solution to the problem of dealing with an overabundance of zero counts.

It postulates that the observed counts arise from a mixture of two latent classes of observations: some structural zeros for whom y_i will always be 0, and the rest, sometimes giving random zeros.

The ZIP model is comprised of two components:

- A model for the binary event of membership in the unobserved (latent) class of those for whom the count is necessarily zero (e.g., “non-publishers”). This is typically taken as a logistic regression for the probability π_i that observation i is in this class, with predictors $z_1, z_2, \ldots, z_q$, giving

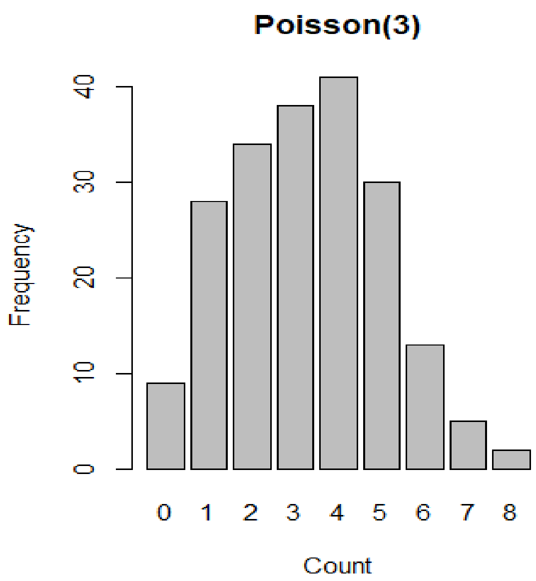
$$logit(\pi_i) = z_i^T \gamma = \gamma_0 + \gamma_1 z_{i1} + \gamma_2 z_{i2} + \cdots + \gamma_q z_{iq}$$

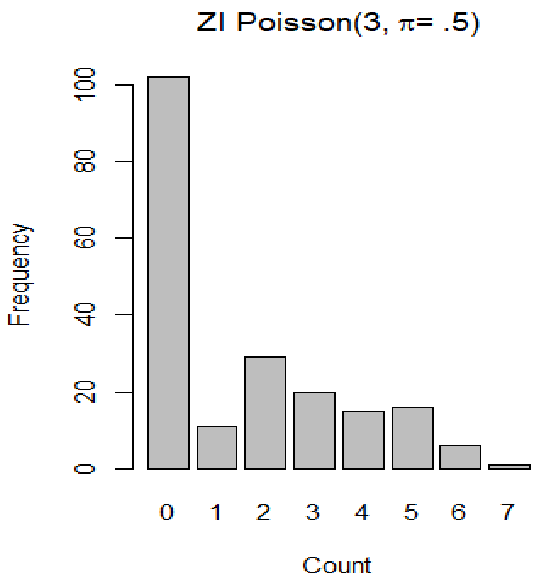
- A Poisson model for the other class (e.g., “publishers”), for whom the observed count may be 0 or positive. This model typically uses the usual log link to predict the mean, using predictors $x_1, x_2, \ldots, x_p$, so
  $log_e \mu(x_i) = x_i^T \beta = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}$.

With this setup, one can show that the probability of observing counts of $y_i = 0$ and $y_i > 0$ are

$$pr(y_i \rfloor x, z) = \pi_i + (1 - \pi_i) e^{-\mu_i}$$
$$pr(y_i \rfloor x, z) = (1 - \pi_i) \times \left[\frac{{\mu_i}^{y_i} e^{-y_i}}{y_i!}\right] \quad , y_i \geq 0.$$

Example





Hurdle Model

A different class of models capable of accounting for excess zero counts is the *hurdle model* (also called the *zero-altered model*).

This model also uses a separate logistic regression sub-model to distinguish counts of y = 0 from larger counts, y > 0. The sub-model for the positive counts is expressed as a (left) *truncated* Poisson or negative-binomial model, excluding the zero counts. As an example, consider a study of behavioural health in which one outcome is the number of cigarettes smoked in one month. All the zero counts will come from non-smokers and smokers will nearly always smoke a positive number.

This differs from the set of ZIP models in that classes of y = 0 and y > 0 are now considered fully observed, rather than latent. Conceptually, there is one process and sub-model accounting for the zero counts and a separate process accounting for the positive counts, once the “hurdle” of y = 0 has been passed. In other words, for ZIP models, the first process generates only extra zeros beyond those of the regular Poisson distribution. For hurdle models, the first process generates all the zeros.

$$pr(y_i = 0 \rfloor x, z) = \pi_i$$

$$pr(y_i \rfloor x, z) = \frac{(1-\pi_i)}{(1-\pi_i)e^{-\mu_i}} \times \left[\frac{{\mu_i}^{y_i} e^{-y_i}}{y_i!}\right] \quad , y_i \geq 0.$$

Limitation of Models

In the literature, hurdle and ZIP models are widely used for analyzing count responses with excessive zeros. However, hurdle and ZIP models do not allow for underdispersion with excessive zeros. In practice, such data sets often exist, for example, the incidence rate of hospitalization, and accident rates when accidents are very rare events, where excess zero counts appear along with underdispersion.

Chapter 3

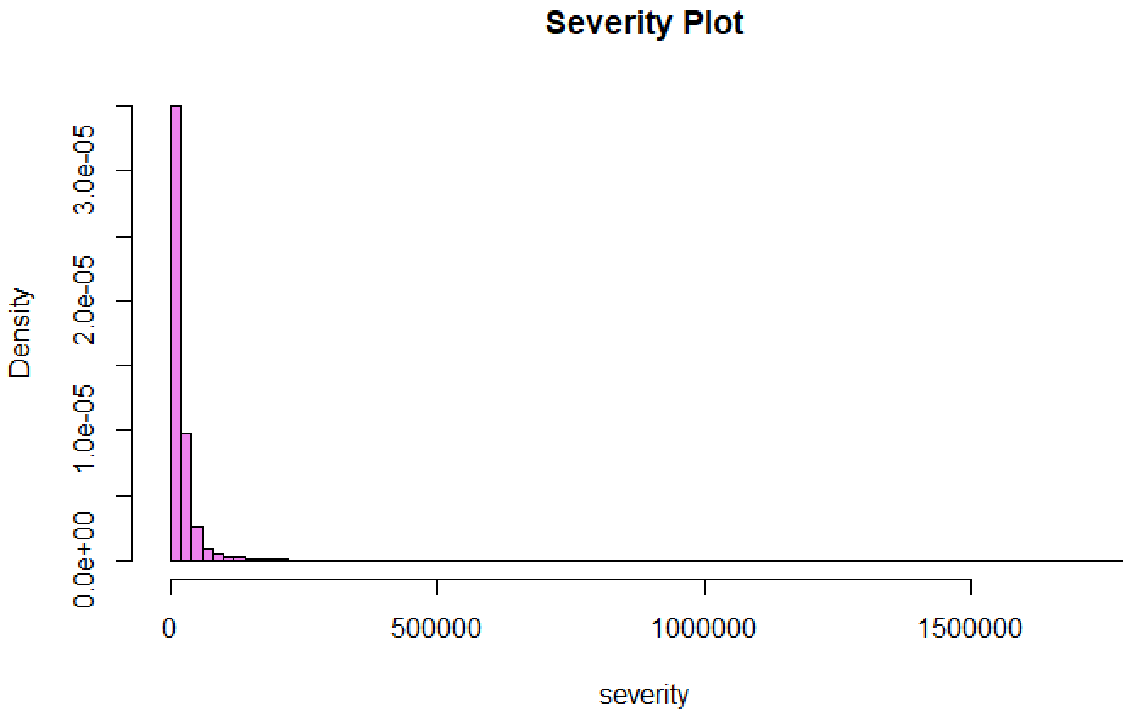
Severity Modelling

Intuition

Once the number of claims is estimated, the claim severity can be modelled: the claim frequency is analyzed, conditionally on the number of claims (which is exactly the exposure).

Severity is the average claim cost. It is needed to calculate the average premium.

Data visualization



Severity Modelling

As the data above has positive skewness, we can consider few similar continuous families of distributions. Those can be Gamma, inverse Gaussian, Lognormal etc.

Here we considered Gamma and Inverse Gaussian.

Generalized Linear Model

As the distribution of the response is again a non-normal one, we have to consider generalised linear model where the response variable may follow the following variables according to the shape of the data.

Gamma Distribution

Gamma is a member of exponential family.

The probability density function of the distribution is given below,

$$f(y;\mu,k) = \frac{\left({}^{k}/_{\mu}\right)^{k}}{\Gamma(k)}\, e^{-{}^{ky}/_{\mu}} y^{k-1} \qquad ,y > 0$$

Inverse Gaussian

Inverse Gaussian is another skewed distribution used fir modeling claim cost.

The probability density function of the distribution is given as follows:

$$f(y;\mu,\lambda_{IG}) = \left\{\frac{\lambda_{IG}}{2\pi y^3}\right\}^{1/2} exp\left\{\frac{-\lambda_{IG}(y-\mu)^2}{2\mu^2 y}\right\} \qquad ,y > 0$$

Goodness of fit measures

The following three criteria are well known and will be useful in selecting among models, with smaller values representing better model fit.

The Akaike information criterion (AIC) is a measure that balances model fit against model simplicity: it takes into account the log likelihood L and the number of parameters *r* (don't forget the scale parameter if it is also estimated). AIC has the form

$$\boldsymbol{AIC} = -2\mathcal{L}+2\mathrm{r}$$

So, it penalizes over fitting - using too much parameters. An alternative form is the corrected AIC given by

$$\boldsymbol{AICC} = -2\mathcal{L} + 2r\frac{n}{n-r-1}$$

where n is the total number of observations used and clearly converges to AIC for large n and small r.

A third, similar measure is the Bayesian information criterion (BIC), which is bigger than the AIC (or AICC) for large enough n :

$$\boldsymbol{BIC} = -2\mathcal{L} + r ln(n)$$

Deviance

Another statistic that is often used to compare models, but has also meaning on its own for a specific model, is the deviance. This is defined as the difference in loglikelihood of the saturated model and the model under consideration:

$$D = 2\varphi[\mathcal{L}(y_i; y_i) - \mathcal{L}(\widehat{\mu_i}; y_i)]$$

Goodness of fit test

LR test

Here we check whether the variables considered in constructing the model is an optimal choice or not. Likelihood Ratio Test is used for the purpose.

Chapter 4

Validation

intuition

A common practice in data science competitions is to iterate over various models to find a better performing model. However, it becomes difficult to distinguish whether this improvement in score is coming because we are capturing the relationship better, or we are just over-fitting the data. To find the right answer for this question, we use validation techniques. This method helps us in achieving more generalized relationships.

Cross Validation Approach

Cross Validation is a technique which involves reserving a particular sample of a dataset on which you do not train the model. Later, you test your model on this sample before finalizing it.

Here are the steps involved in cross validation:

1. You reserve a sample data set

2. Train the model using the remaining part of the dataset

3. Use the reserve sample of the test (validation) set. This will help you in gauging the effectiveness of your model's performance. If your model delivers a positive result on validation data, go ahead with the current model.

There are various methods available for performing cross validation. Few of them are validation set approach, Leave one out cross validation (LOOCV), k-fold cross validation, Stratified k-fold cross validation, Adversarial Validation, Cross Validation for time series, Custom Cross Validation Techniques.

The validation set approach

In this approach, we reserve 50% of the dataset for validation and the remaining 50% for model training. However, a major disadvantage of this approach is that since we are training a model on only 50% of the dataset, there is a huge possibility that we might miss out on some interesting information about the data which will lead to a higher bias.

Leave one out cross validation (LOOCV)

In this approach, we reserve only one data point from the available dataset, and train the model on the rest of the data. This process iterates for each data point. This also has its own advantages and disadvantages.

- We make use of all data points, hence the bias will be low

- We repeat the cross validation process n times (where n is number of data points) which results in a higher execution time

- This approach leads to higher variation in testing model effectiveness because we test against one data point. So, our estimation gets highly influenced by the data point. If the data point turns out to be an outlier, it can lead to a higher variation

Frequency Model validation

The (minimal) frequency at which a model has to be re-validated. This is often determined in the model validation policy of the bank.

The model validation frequency typically depends on the amount of model risk that is associated with the model. As an example, a model that is used extensively might have to be validated yearly while a model that is only used sparsely might have to be revalidated only once every three years. Typically, the amount of model risk that is carried by the model is expressed in terms of model risk tiers.

The main objective of the methodology presented is to validate a model on the frequency domain. To this end a time domain validation procedure based on testing the residual whiteness is modified to achieve the pursued objectives. The idea is as follows. It is assumed that if the residual is white noise the model is validated because the residual contains no further useful information that could be used to improve the model accuracy. This test is usually performed in the time domain by studying the residual autocorrelation, the number of sign changes, etc

We translate the time domain residual to the frequency domain by its discrete Fourier transform. Moreover, the statistical properties of the spectrum of a white noise signal are calculated. The objective is to test if the spectrum calculated from the residual has properties of white noise. As a result, one unique test in the time domain has been translated to N different tests in the frequency domain. We check if the k th frequency component of the spectrum has the properties of a typical frequency component of a white noise.

In the affirmative case we have no reason to believe that the model is invalid on that frequency component. On the other hand, if there are certain frequency components that clearly do not behave accordingly with the statistical properties of white noise then it is likely that at this frequency range there is an important mismatch between the model and the plant. As a result the model is invalid for that frequency range.

Chapter 5

Results & Discussion

Data

This dataset contains data from a certain insurance company which will remain anonymous. The data represent car insurance policies (claims and costs with respect to third party liability only), spread over several years (Not specified). It deals with portfolio of policies where different type policies are involved.

Variable Overview

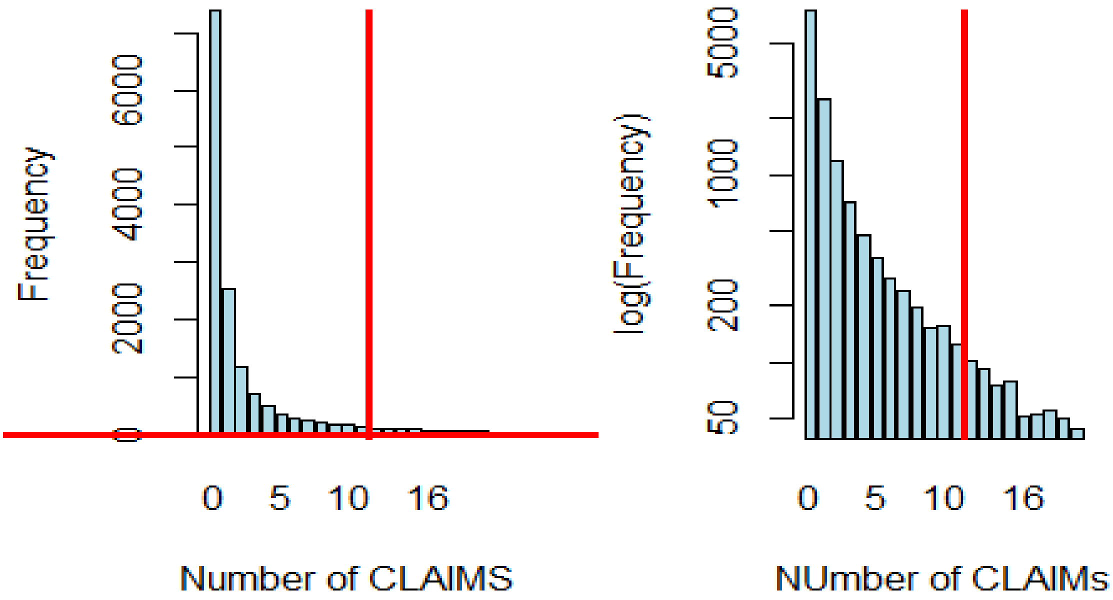
Variable Name	Type	Levels	Description
POLICY_YEAR	Numeric	1	Issue of policy in 2016
CC_BAND_GROUP	Categorical	1001 to 1300 1301 to 1500 1501 to 2000 1<1000	Engine Capacity, measured in Cubic Capacity (CC). Higher capacity of engine pays Higher premium
AGE_FLOOR2	Categorical	1 2 3 4 5 <1	Age of the vehicle
FUEL_TYPE_GROUP	Categorical	Diesel & Others 1Petrol	Type of Fuel used in vehicle
NCB	Categorical	20 25 35 45 1<20 50 65 70 0	NO CLAIM BONUS. Customer gets bonus by not making any claim in a particular year by means of discount of the following year.
ZERO_DEP_FLAG	Categorical	YES NO	**Zero depreciation** costs anywhere between 15-20% of the standard premium and is a MUST BUY for all new or

			relatively new (up to 5 years) cars. **Zero depreciation** car insurance proves to be beneficial to: People with new cars. People with luxury cars.
IDV_BAND	Categorical	3to5Lakh 5to10Lakh 10to15Lakh 15to25Lakh 1<3Lakh 25to50Lakh more_than_50Lakh	INSURED DECLARED VALUE. Maximum amount for which the Car is insured
WRITTEN_PREMIUM	Numerical		Premium Insurer will get at the commencement of new policy and renewal of existing policy. It is different from total Premium gained as some policies can dis continue.
POLICY_COUNT	Numerical		No of Policy in that particular portfolio.
IDV	Numerical		In general it is the actual value of the vehicle. It is calculated every year as with time, vehicle's value depreciates. It is ex-showroom price/current market price minus depreciation on its parts.
CLAIM_COUNT	Numerical		
INCURRED_CLAIM	Numerical		Insured event happened & for which the insurer is liable to pay if claim is made.
VEHICLE_MAKE	Categorical	Make B Make C Make D Make E Make A Others	Type engine used (or, model of the vehicle)
AVG_IDV	Numerical		Throughout the year the average value of

			the Vehicle declared by insured after adjusting for the depreciation on the parts
FUEL_AGE_NILDEP	Categorical		Combination of the following variables -Fuel Type -Age of vehicle -Depriciation applied on the parts or not.
NCB_ADJ_POLICY_COUNT	Numerical		Policy Count Adjusted after the considering the policies at that discount level currently
SEVERITY_BY_IDV	Continuous		Average Claim cost per Unit Insurance Cover.

Summary of the Variables

- Histogram of CLAIM DATA and transformed data



The frequencies of 0–2 articles account for over 75% of the total, so that the frequencies of the larger counts get lost in the display. To accommodate the zero frequencies, the plot shows *log(Frequency+1)*, avoiding errors from *log(0)*. It can be seen that log frequency decreases steadily up to 15 Claims and then levels off approximately.

The vertical bar shows the mean of the dataset and horizontal lines show mean ±1 standard deviation.

Mean Variance Relation

mean	var	ratio
14.57892	13501.94843	926.12791

Summary Analysis of Independent Variable

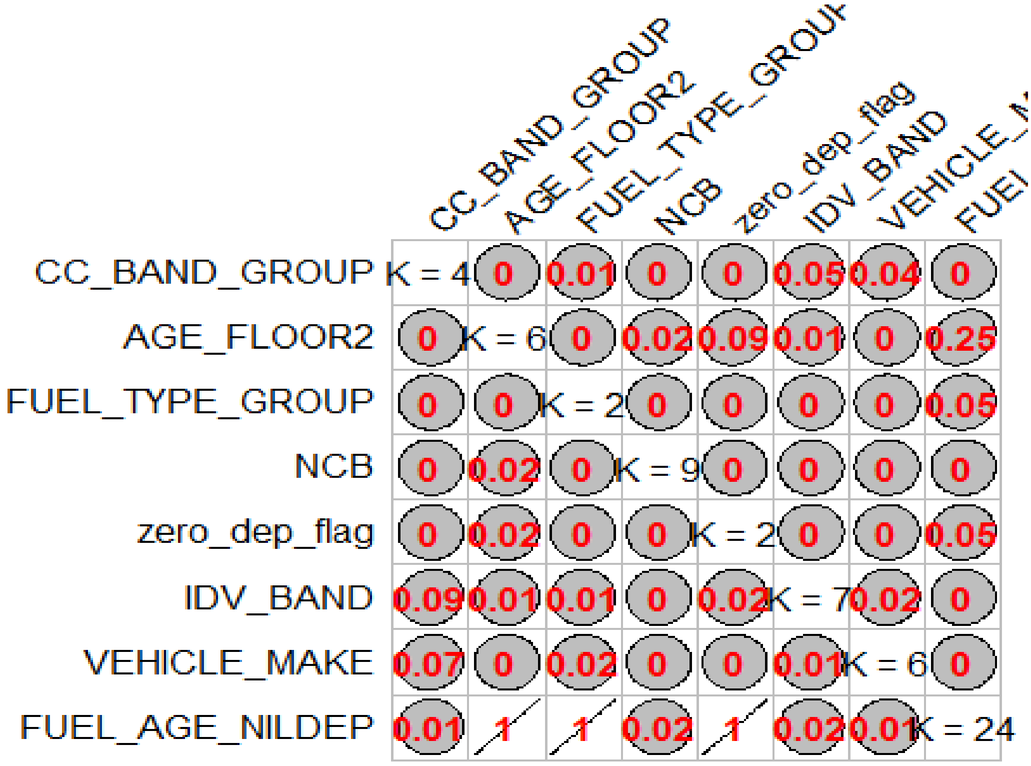
Variable Selection Method

The variables here considered are both continuous and categorical. Basic correlation of continuous variables can be considered for basic model selection. But for categorical variables it is not so easy. We will make Stepwise selection method in Generalized Linear Model.

Hypothetically the following categorical variables may be more influential for modelling the claim count

Car Capacity
Age of Car
Fuel Type
Make of the Vehicle
Depreciation

Dependencies between the explanatory variables

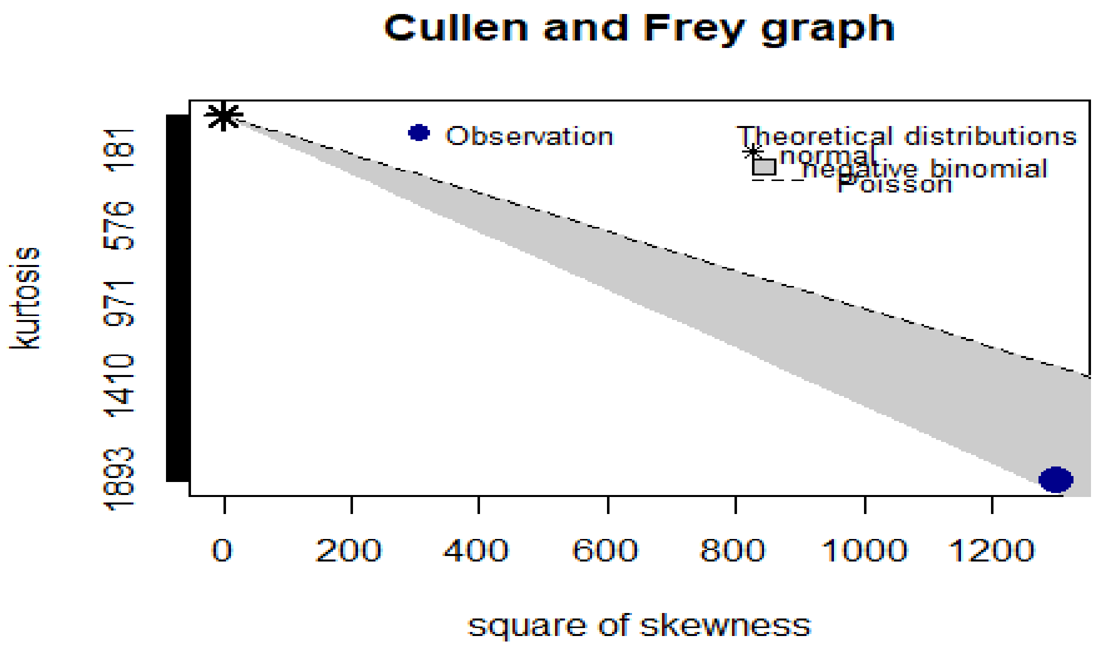


Goodmann Kruskal Gamma

Here the assumption independence of explanatory variable is checked. The k values are the number of levels of the respective variables. The measure is considered here is Goodmann kruskal Gamma, similar to the correlation coefficient of the continuous data.

It is seen that the variables are almost independent among themselves. The last variable is the combination of few variables, so it has a prefect positive relation with respective variable.

Distribution




```
 min:  0   max:  7214 
median:  1 
mean:  14.57892 
estimated sd:  116.1979 
estimated skewness:  36.05447 
estimated kurtosis:  1892.395
```

Here based on the Kurtosis and Skewness we can see that the data fits in between the Negative Binomial and Poisson distribution.

Fitting a Basic GLM model on full variables set using Poisson and revised model.

```
AIC(obj_pois_full)

## 528020.4

AIC(obj1)

## 60207.28
```

Here We can see the full model is not so good as AIC value is too high. Rather the reduced model gives less AIC Score.

Comparing the both models to check which model is the better one

Here we are using ‘lmtest’ will compare between full model and the selected variables’ model.

	Df	Log-Likelihood	Df	Chi-square	Pr(>Chisq)
Model 1	52	-263958			
Model 2	16	-30088	-36	467741	< 2.2e-16 ***

Significance. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

We can see that the lmtest is indicating Model 2 as significant one. Now we got our optimal model.

Model 2: ***CLAIM_COUNT ~ CC_BAND_GROUP + AGE_FLOOR2 + FUEL_TYPE_GROUP + VEHICLE_MAKE + zero_dep_flag +*** *offset (log* ***(NCB_ADJ_POLICY_COUNT))***

Further we can see that if we consider the combination of variables (Fuel Type, Age of vehicle, Depreciation), it will be a better fit as the combination makes more relevant portfolio wise division of the policies.

RESULT

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-2.077934	0.027362	-75.943	< 2e-16	***
CC_BAND_GROUP1001 to 1300	0.329118	0.005532	59.499	< 2e-16	***
CC_BAND_GROUP1301 to 1500	0.380094	0.008915	42.633	< 2e-16	***
CC_BAND_GROUP1501 to 2000	0.359101	0.010291	34.896	< 2e-16	***
VEHICLE_MAKEMakeB	-0.109752	0.013161	-8.339	< 2e-16	***
VEHICLE_MAKEMakeC	-0.286658	0.012742	-22.496	< 2e-16	***
VEHICLE_MAKEMakeD	-0.223077	0.011958	-18.655	< 2e-16	***
VEHICLE_MAKEMakeE	-0.048515	0.005683	-8.536	< 2e-16	***
VEHICLE_MAKEOthers	-0.161116	0.007199	-22.380	< 2e-16	***
FUEL_AGE_NILDEP1Petrol<1YES	0.829751	0.027915	29.724	< 2e-16	***
FUEL_AGE_NILDEP1Petrol1NO	-0.041110	0.032947	-1.248	0.212	
FUEL_AGE_NILDEP1Petrol1YES	0.659793	0.028146	23.442	< 2e-16	***
FUEL_AGE_NILDEP1Petrol2NO	-0.019907	0.030752	-0.647	0.517	
FUEL_AGE_NILDEP1Petrol2YES	0.678451	0.028456	23.842	< 2e-16	***
FUEL_AGE_NILDEP1Petrol3NO	0.042905	0.029672	1.446	0.148	
FUEL_AGE_NILDEP1Petrol3YES	0.749829	0.029143	25.729	< 2e-16	***
FUEL_AGE_NILDEP1Petrol4NO	-0.010668	0.029282	-0.364	0.716	
FUEL_AGE_NILDEP1Petrol4YES	0.865585	0.029915	28.935	< 2e-16	***
FUEL_AGE_NILDEP1Petrol5NO	-0.289202	0.027597	-10.479	< 2e-16	***
FUEL_AGE_NILDEP1Petrol5YES	0.589279	0.071750	8.213	< 2e-16	***
FUEL_AGE_NILDEPDiesel&Others<1NO	0.306072	0.052723	5.805	6.42e-09	***
FUEL_AGE_NILDEPDiesel&Others<1YES	0.968828	0.028367	34.153	< 2e-16	***
FUEL_AGE_NILDEPDiesel&Others1NO	0.245350	0.040145	6.112	9.86e-10	***
FUEL_AGE_NILDEPDiesel&Others1YES	0.930630	0.028568	32.576	< 2e-16	***
FUEL_AGE_NILDEPDiesel&Others2NO	0.326093	0.034002	9.590	< 2e-16	***
FUEL_AGE_NILDEPDiesel&Others2YES	0.939363	0.028784	32.635	< 2e-16	***
FUEL_AGE_NILDEPDiesel&Others3NO	0.290365	0.031955	9.087	< 2e-16	***
FUEL_AGE_NILDEPDiesel&Others3YES	0.958994	0.029216	32.824	< 2e-16	***
FUEL_AGE_NILDEPDiesel&Others4NO	0.319046	0.031250	10.209	< 2e-16	***
FUEL_AGE_NILDEPDiesel&Others4YES	1.058205	0.030104	35.151	< 2e-16	***
FUEL_AGE_NILDEPDiesel&Others5NO	0.171568	0.029191	5.878	4.16e-09	***
FUEL_AGE_NILDEPDiesel&Others5YES	0.881258	0.093570	9.418	< 2e-16	***

AIC value of Individual and Combination of the variables.

Individual	Combination
60207.28	59625.34

There is a significant Difference in our goodness of fit measure. So, we can consider the following variable set, with **offset** term log (**NCB_ADJ_POL_COUNT**).

Here we are taking NCB_ADJ_POLICY Count instead of Exposure, as the Exposure is the not available for this portfolio type data. So, we can put the policy count as the weighting term for our Claim Count, as it will balance the effect of the high claim frequency with high policy count for a particular portfolio.

CLAIM_COUNT~CC_BAND_GROUP+VEHICLE_MAKE+FUEL_AGE_NILDEP

Negative Binomial Distribution

As we know that the data is over-dispersed, Poisson may not be a good choice as it's mean and variance are equal. But for negative Binomial distribution the data suits more well as the variance of the data is more than the mean.

```
Coefficients:
                                    Estimate Std. Error z value Pr(>|z|)
(Intercept)                          -2.005881   0.045129 -44.448  < 2e-16 ***
CC_BAND_GROUP1001 to 1300            0.215205   0.015485  13.898  < 2e-16 ***
CC_BAND_GROUP1301 to 1500            0.292802   0.018935  15.464  < 2e-16 ***
CC_BAND_GROUP1501 to 2000            0.274285   0.019790  13.860  < 2e-16 ***
VEHICLE_MAKEMakeB                    -0.117838   0.021747  -5.419 6.01e-08 ***
VEHICLE_MAKEMakeC                    -0.236242   0.024953  -9.468  < 2e-16 ***
VEHICLE_MAKEMakeD                    -0.229165   0.024316  -9.424  < 2e-16 ***
VEHICLE_MAKEMakeE                    0.043735   0.015800   2.768 0.005640 **
VEHICLE_MAKEOthers                   -0.106051   0.016508  -6.424 1.33e-10 ***
FUEL_AGE_NILDEP1Petrol<1YES          0.594626   0.050529  11.768  < 2e-16 ***
FUEL_AGE_NILDEP1Petrol1NO            0.004302   0.055005   0.078 0.937658
FUEL_AGE_NILDEP1Petrol1YES           0.642648   0.048939  13.132  < 2e-16 ***
FUEL_AGE_NILDEP1Petrol2NO            0.031544   0.052174   0.605 0.545448
FUEL_AGE_NILDEP1Petrol2YES           0.682679   0.048808  13.987  < 2e-16 ***
FUEL_AGE_NILDEP1Petrol3NO            0.091929   0.050453   1.822 0.068445 .
FUEL_AGE_NILDEP1Petrol3YES           0.741483   0.049213  15.067  < 2e-16 ***
FUEL_AGE_NILDEP1Petrol4NO            0.075008   0.049435   1.517 0.129192
FUEL_AGE_NILDEP1Petrol4YES           0.854847   0.049857  17.146  < 2e-16 ***
FUEL_AGE_NILDEP1Petrol5NO            -0.136161   0.046736  -2.913 0.003576 **
FUEL_AGE_NILDEP1Petrol5YES           0.576014   0.086084   6.691 2.21e-11 ***
FUEL_AGE_NILDEPDiesel&Others<1NO   0.292826   0.072947   4.014 5.96e-05 ***
FUEL_AGE_NILDEPDiesel&Others<1YES  0.801574   0.052455  15.281  < 2e-16 ***
FUEL_AGE_NILDEPDiesel&Others1NO    0.236378   0.060863   3.884 0.000103 ***
FUEL_AGE_NILDEPDiesel&Others1YES   0.884777   0.050273  17.599  < 2e-16 ***
FUEL_AGE_NILDEPDiesel&Others2NO    0.304801   0.054681   5.574 2.49e-08 ***
FUEL_AGE_NILDEPDiesel&Others2YES   0.914969   0.049711  18.406  < 2e-16 ***
FUEL_AGE_NILDEPDiesel&Others3NO    0.243463   0.052597   4.629 3.68e-06 ***
FUEL_AGE_NILDEPDiesel&Others3YES   0.939971   0.049821  18.867  < 2e-16 ***
FUEL_AGE_NILDEPDiesel&Others4NO    0.270917   0.051791   5.231 1.69e-07 ***
FUEL_AGE_NILDEPDiesel&Others4YES   1.015629   0.051014  19.909  < 2e-16 ***
FUEL_AGE_NILDEPDiesel&Others5NO    0.156573   0.049689   3.151 0.001627 **
FUEL_AGE_NILDEPDiesel&Others5YES   0.844925   0.105938   7.976 1.52e-15 ***
---
```

Comparison is based on AIC value of the model.

Poisson	Negative Binomial
59625.34	48286.51

Some Special Distributions

In addition to overdispersion, many sets of empirical data exhibit a greater prevalence of zero counts than can be accommodated by the Poisson or negative-binomial models.

If we take a sample from data to consider the relative frequency distribution, we can see the following.

Claim Count	Proportion
0	0.474
1	0.170
2	0.073
3	0.043
4	0.029

We saw this in the CLAIM Count data set, where there were many policy holders whose claims cannot be processed due to mandatory minimum policy period for Claim application.

Similarly, the distribution of insurance claims often shows large numbers who make no claims because of under-reporting of small claims, policy deductible provisions, and desire to avoid premium increases.

One reasonable form of explanation is that the observed zero counts reflect a mixture of the two latent classes—those who simply have not claimed for any accident and those whose claims cannot be considered because of some constraint of policy.

Zero-inflated models

Zero Inflated distributions are used for count data with more number of zeros. Here both types of zeros are used like structural and actual zeros.

Here we considered both Poisson and Negative Binomial distribution for the modelling part of the structured zero along with positive count.

Bar Plot of the Poisson and Zero Inflated Poisson distribution

Model Output

Count model coefficients (negbin with log link):

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	2.11173	0.12756	16.555	< 2e-16	***
CC_BAND_GROUP1001 to 1300	-0.02568	0.05656	-0.454	0.6497	
CC_BAND_GROUP1301 to 1500	-0.06317	0.07524	-0.840	0.4011	
CC_BAND_GROUP1501 to 2000	-1.01509	0.06905	-14.700	< 2e-16	***
VEHICLE_MAKEMakeB	-2.43767	0.07150	-34.094	< 2e-16	***
VEHICLE_MAKEMakeC	-2.05394	0.08872	-23.150	< 2e-16	***
VEHICLE_MAKEMakeD	-2.14252	0.08442	-25.378	< 2e-16	***
VEHICLE_MAKEMakeE	-0.88355	0.06272	-14.088	< 2e-16	***
VEHICLE_MAKEOthers	-1.44615	0.06538	-22.120	< 2e-16	***
FUEL_AGE_NILDEP1Petrol<1YES	2.37834	0.14981	15.876	< 2e-16	***
FUEL_AGE_NILDEP1Petrol1NO	0.31601	0.15201	2.079	0.0376	*
FUEL_AGE_NILDEP1Petrol1YES	2.13453	0.14577	14.643	< 2e-16	***
FUEL_AGE_NILDEP1Petrol2NO	0.61595	0.14508	4.246	2.18e-05	***
FUEL_AGE_NILDEP1Petrol2YES	1.81703	0.14290	12.716	< 2e-16	***
FUEL_AGE_NILDEP1Petrol3NO	0.88542	0.14234	6.221	4.95e-10	***
FUEL_AGE_NILDEP1Petrol3YES	1.49666	0.14279	10.482	< 2e-16	***
FUEL_AGE_NILDEP1Petrol4NO	1.02707	0.13885	7.397	1.39e-13	***
FUEL_AGE_NILDEP1Petrol4YES	1.42016	0.14595	9.731	< 2e-16	***
FUEL_AGE_NILDEP1Petrol5NO	2.45026	0.13276	18.457	< 2e-16	***
FUEL_AGE_NILDEP1Petrol5YES	-0.80416	0.20584	-3.907	9.36e-05	***
FUEL_AGE_NILDEPDiesel&Others<1NO	-0.14556	0.18630	-0.781	0.4346	

```
FUEL_AGE_NILDEPDiesel&Others<1YES  2.58903    0.15941   16.241  < 2e-16 ***
FUEL_AGE_NILDEPDiesel&Others1NO    0.29934    0.16874    1.774   0.0761 .
FUEL_AGE_NILDEPDiesel&Others1YES   2.55624    0.15396   16.603  < 2e-16 ***
FUEL_AGE_NILDEPDiesel&Others2NO    0.72598    0.15421    4.708 2.51e-06 ***
FUEL_AGE_NILDEPDiesel&Others2YES   2.40391    0.15003   16.022  < 2e-16 ***
FUEL_AGE_NILDEPDiesel&Others3NO    0.93133    0.14993    6.212 5.24e-10 ***
FUEL_AGE_NILDEPDiesel&Others3YES   2.16885    0.14886   14.569  < 2e-16 ***
FUEL_AGE_NILDEPDiesel&Others4NO    1.09430    0.14911    7.339 2.15e-13 ***
FUEL_AGE_NILDEPDiesel&Others4YES   1.89222    0.15246   12.411  < 2e-16 ***
FUEL_AGE_NILDEPDiesel&Others5NO    1.73902    0.14454   12.031  < 2e-16 ***
FUEL_AGE_NILDEPDiesel&Others5YES  -0.57097    0.24828   -2.300   0.0215 *
Log(theta)                        -1.61198    0.01343 -120.069  < 2e-16 ***

Zero-inflation model coefficients (binomial with logit link):
                                  Estimate Std. Error z value Pr(>|z|)
(Intercept)                        -6.99226    4.06966  -1.718   0.0858 .
CC_BAND_GROUP1001 to 1300          -4.74037    9.32572  -0.508   0.6112
CC_BAND_GROUP1301 to 1500           6.63537    4.01273   1.654   0.0982 .
CC_BAND_GROUP1501 to 2000           6.83007    4.04252   1.690   0.0911 .
VEHICLE_MAKEMakeB                  -9.64221         NA      NA       NA
VEHICLE_MAKEMakeC                  -8.83190    4.75316  -1.858   0.0632 .
VEHICLE_MAKEMakeD                  -9.40431    4.91701  -1.913   0.0558 .
VEHICLE_MAKEMakeE                 -11.84328   13.56921  -0.873   0.3828
VEHICLE_MAKEOthers                -11.43703   11.28845  -1.013   0.3110
FUEL_AGE_NILDEP1Petrol<1YES        -0.61558    1.10140  -0.559   0.5762
FUEL_AGE_NILDEP1Petrol1NO          -0.81843    1.23645  -0.662   0.5080
FUEL_AGE_NILDEP1Petrol1YES         -0.52376    1.05635  -0.496   0.6200
FUEL_AGE_NILDEP1Petrol2NO          -0.65184    1.10207  -0.591   0.5542
FUEL_AGE_NILDEP1Petrol2YES         -0.89402    1.09025  -0.820   0.4122
FUEL_AGE_NILDEP1Petrol3NO          -0.22847    1.01041  -0.226   0.8211
FUEL_AGE_NILDEP1Petrol3YES         -0.67827    1.08057  -0.628   0.5302
FUEL_AGE_NILDEP1Petrol4NO           0.11808    1.00108   0.118   0.9061
FUEL_AGE_NILDEP1Petrol4YES          0.35666    1.05276   0.339   0.7348
FUEL_AGE_NILDEP1Petrol5NO          -1.17227    1.20296  -0.974   0.3298
FUEL_AGE_NILDEP1Petrol5YES         -5.12077   19.97094  -0.256   0.7976
FUEL_AGE_NILDEPDiesel&Others<1NO   -4.02846         NA      NA       NA
FUEL_AGE_NILDEPDiesel&Others<1YES  -0.28325    1.12500  -0.252   0.8012
FUEL_AGE_NILDEPDiesel&Others1NO     5.04572    3.63649   1.388   0.1653
FUEL_AGE_NILDEPDiesel&Others1YES    0.24493    1.11741   0.219   0.8265
FUEL_AGE_NILDEPDiesel&Others2NO     0.94462    1.66119   0.569   0.5696
FUEL_AGE_NILDEPDiesel&Others2YES    1.03617    1.17939   0.879   0.3796
FUEL_AGE_NILDEPDiesel&Others3NO     1.06195    1.31080   0.810   0.4179
FUEL_AGE_NILDEPDiesel&Others3YES    2.37800    1.39972   1.699   0.0893 .
FUEL_AGE_NILDEPDiesel&Others4NO     2.11177    1.42963   1.477   0.1396
FUEL_AGE_NILDEPDiesel&Others4YES   -0.07731    1.32543  -0.058   0.9535
FUEL_AGE_NILDEPDiesel&Others5NO     0.37796    1.08905   0.347   0.7286
FUEL_AGE_NILDEPDiesel&Others5YES    2.18322    4.64560   0.470   0.6384
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Theta = 0.1995
```

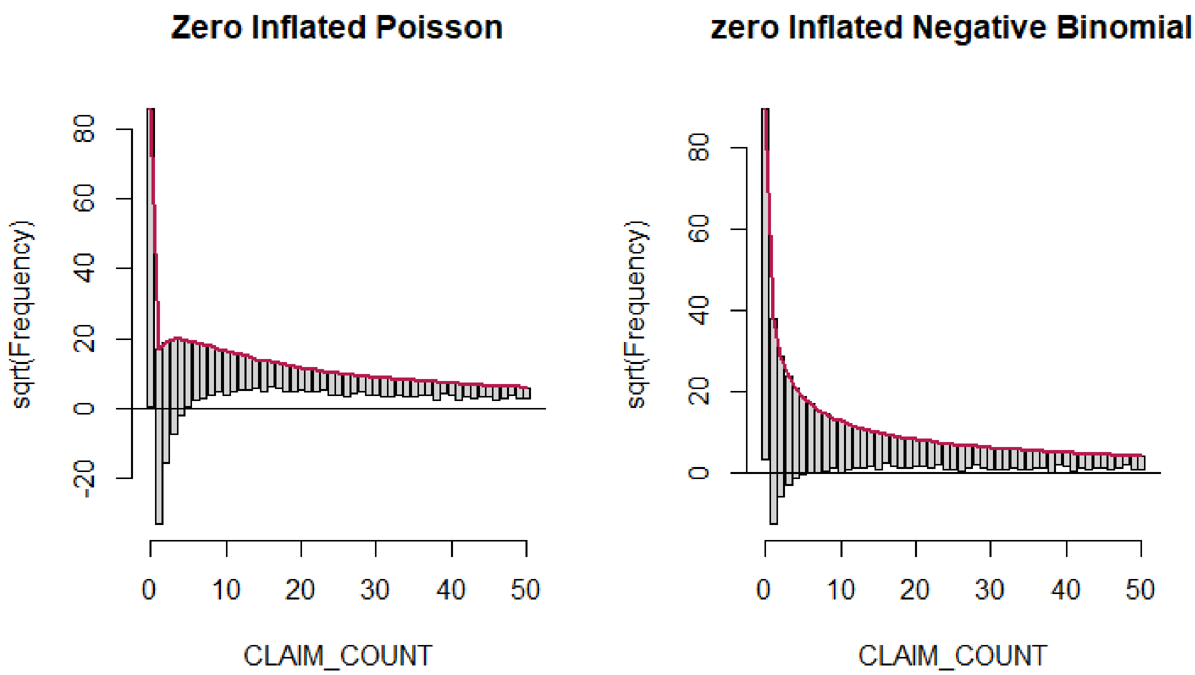
Testing which model is better using “voung test”

	Vuong z-statistic	Hypothesis	p-value
Raw	-35.63566	model2 > model1	< 2.22e-16
AIC-corrected	-35.63566	model2 > model1	< 2.22e-16
BIC-corrected	-35.63566	model2 > model1	< 2.22e-16

Voung test is considered to test the goodness of fit for both the models. We can see that the model 2, i.e. Negative Binomial better.

Visualization of Goodness of Fit

The tops of the bars are the *expected frequencies* of the counts given the model. The counts are plotted on the square-root scale to help visualize smaller frequencies. The red line shows the fitted frequencies as a smooth curve. The x-axis is actually a horizontal reference line. Bars that hang below the line show underfitting, bars that hang above show overfitting. In this case it's hard to see any over or underfitting because we fit the right model. In a moment we'll see some rootograms that clearly identify an ill-fitting model.



Though the Zero Inflated Poisson model accurately estimated the zero value, it underestimated the less frequent values. Whereas, Negative Binomial is estimating quite better than Poisson in terms of rest of the values except the overestimating Zero and underestimating 1-5 values.

Hurdle Model

We know Hurdle models are used for making the two different modeling for two parts of the data. Binomial distribution is fitted for the hurdle model & other distribution for the positive count data (more than zero). Here two distributions are modeled as independently.

This differs from the set of ZIP models in that classes of y = 0 and y > 0 are now considered fully observed, rather than latent.

Model Output

```
Count model coefficients (truncated negbin with log link):
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)               -8.28902   11.33094  -0.732 0.464450
CC_BAND_GROUP1001 to 1300      -0.13318    0.08033  -1.658 0.097315 .
CC_BAND_GROUP1301 to 1500      -0.16045    0.11440  -1.403 0.160746
CC_BAND_GROUP1501 to 2000      -1.14922    0.10130 -11.345  < 2e-16 ***
VEHICLE_MAKEMakeB           -2.81296    0.09863 -28.521  < 2e-16 ***
VEHICLE_MAKEMakeC           -2.18031    0.13210 -16.505  < 2e-16 ***
VEHICLE_MAKEMakeD           -2.26203    0.12377 -18.276  < 2e-16 ***
VEHICLE_MAKEMakeE           -1.00902    0.08767 -11.509  < 2e-16 ***
VEHICLE_MAKEOthers         -1.63136    0.09305 -17.533  < 2e-16 ***
FUEL_AGE_NILDEP1Petrol<1YES         2.72050    0.21890  12.428  < 2e-16 ***
FUEL_AGE_NILDEP1Petrol1NO         0.31923    0.21422   1.490 0.136183
FUEL_AGE_NILDEP1Petrol1YES         2.31194    0.20638  11.203  < 2e-16 ***
FUEL_AGE_NILDEP1Petrol2NO         0.59577    0.20309   2.934 0.003351 **
FUEL_AGE_NILDEP1Petrol2YES         1.89743    0.19995   9.489  < 2e-16 ***
FUEL_AGE_NILDEP1Petrol3NO         0.92718    0.20037   4.627 3.70e-06 ***
```

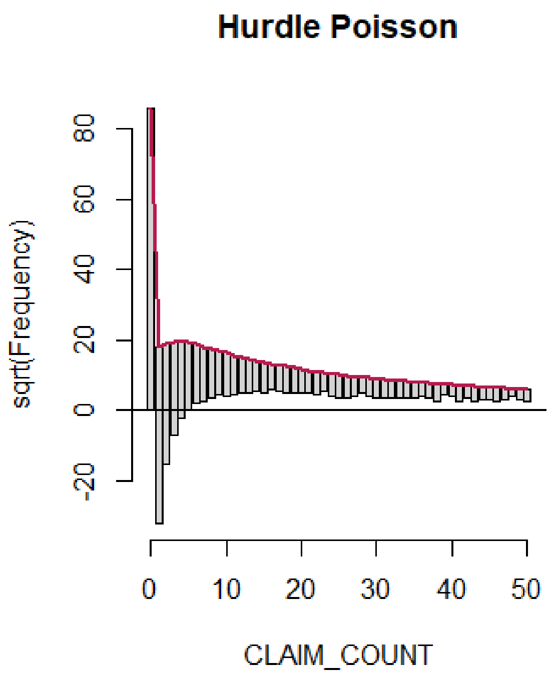
```
FUEL_AGE_NILDEP1Petrol3YES         1.57778    0.20017   7.882 3.22e-15 ***
FUEL_AGE_NILDEP1Petrol4NO          1.06808    0.19522   5.471 4.47e-08 ***
FUEL_AGE_NILDEP1Petrol4YES         1.39636    0.20221   6.905 5.01e-12 ***
FUEL_AGE_NILDEP1Petrol5NO          2.73865    0.18992  14.420  < 2e-16 ***
FUEL_AGE_NILDEP1Petrol5YES        -1.52208    0.27970  -5.442 5.28e-08 ***
FUEL_AGE_NILDEPDiesel&Others<1NO   -0.29522    0.25747  -1.147 0.251533
FUEL_AGE_NILDEPDiesel&Others<1YES   2.92161    0.23199  12.593  < 2e-16 ***
FUEL_AGE_NILDEPDiesel&Others1NO     0.25056    0.23476   1.067 0.285842
FUEL_AGE_NILDEPDiesel&Others1YES    2.79541    0.21968  12.725  < 2e-16 ***
FUEL_AGE_NILDEPDiesel&Others2NO     0.82674    0.21860   3.782 0.000156 ***
FUEL_AGE_NILDEPDiesel&Others2YES    2.60851    0.21265  12.266  < 2e-16 ***
FUEL_AGE_NILDEPDiesel&Others3NO     1.06239    0.21299   4.988 6.10e-07 ***
FUEL_AGE_NILDEPDiesel&Others3YES    2.32814    0.20966  11.104  < 2e-16 ***
FUEL_AGE_NILDEPDiesel&Others4NO     1.25731    0.21229   5.923 3.17e-09 ***
FUEL_AGE_NILDEPDiesel&Others4YES    1.99888    0.21319   9.376  < 2e-16 ***
FUEL_AGE_NILDEPDiesel&Others5NO     1.98399    0.20758   9.558  < 2e-16 ***
FUEL_AGE_NILDEPDiesel&Others5YES   -1.32612    0.32978  -4.021 5.79e-05 ***
Log(theta)                       -12.48502   11.32961  -1.102 0.270470
Zero hurdle model coefficients (binomial with logit link):
                                  Estimate Std. Error z value Pr(>|z|)
(Intercept)                       -0.25553    0.11109  -2.300 0.021437 *
CC_BAND_GROUP1001 to 1300          0.11539    0.04922   2.344 0.019066 *
CC_BAND_GROUP1301 to 1500         -0.09809    0.05351  -1.833 0.066793 .
CC_BAND_GROUP1501 to 2000         -0.46792    0.05390  -8.681  < 2e-16 ***
VEHICLE_MAKEMakeB                 -0.48324    0.06021  -8.027 1.00e-15 ***
VEHICLE_MAKEMakeC                 -0.45787    0.06710  -6.824 8.88e-12 ***
VEHICLE_MAKEMakeD                 -0.52302    0.06600  -7.924 2.30e-15 ***
VEHICLE_MAKEMakeE                  0.13992    0.05302   2.639 0.008310 **
VEHICLE_MAKEOthers                -0.06662    0.05190  -1.284 0.199269
FUEL_AGE_NILDEP1Petrol<1YES        0.60508    0.13113   4.614 3.95e-06 ***
FUEL_AGE_NILDEP1Petrol1NO          0.19594    0.13139   1.491 0.135878
FUEL_AGE_NILDEP1Petrol1YES         0.87892    0.12804   6.864 6.67e-12 ***
FUEL_AGE_NILDEP1Petrol2NO          0.37701    0.12543   3.006 0.002649 **
FUEL_AGE_NILDEP1Petrol2YES         0.93730    0.12512   7.491 6.83e-14 ***
FUEL_AGE_NILDEP1Petrol3NO          0.40958    0.12304   3.329 0.000872 ***
FUEL_AGE_NILDEP1Petrol3YES         0.77004    0.12395   6.213 5.21e-10 ***
FUEL_AGE_NILDEP1Petrol4NO          0.49728    0.11999   4.144 3.41e-05 ***
FUEL_AGE_NILDEP1Petrol4YES         0.89774    0.12637   7.104 1.21e-12 ***
FUEL_AGE_NILDEP1Petrol5NO          0.88112    0.11567   7.618 2.58e-14 ***
FUEL_AGE_NILDEP1Petrol5YES        -0.02285    0.16709  -0.137 0.891216
FUEL_AGE_NILDEPDiesel&Others<1NO   0.09150    0.15751   0.581 0.561279
FUEL_AGE_NILDEPDiesel&Others<1YES  0.80175    0.13867   5.782 7.40e-09 ***
FUEL_AGE_NILDEPDiesel&Others1NO    0.21427    0.14284   1.500 0.133607
FUEL_AGE_NILDEPDiesel&Others1YES   0.99217    0.13434   7.385 1.52e-13 ***
FUEL_AGE_NILDEPDiesel&Others2NO    0.33456    0.13235   2.528 0.011474 *
FUEL_AGE_NILDEPDiesel&Others2YES   1.04056    0.13118   7.932 2.15e-15 ***
FUEL_AGE_NILDEPDiesel&Others3NO    0.41215    0.12862   3.204 0.001353 **
FUEL_AGE_NILDEPDiesel&Others3YES   1.00686    0.12946   7.777 7.41e-15 ***
FUEL_AGE_NILDEPDiesel&Others4NO    0.43926    0.12800   3.432 0.000600 ***
FUEL_AGE_NILDEPDiesel&Others4YES   0.99402    0.13260   7.496 6.56e-14 ***
FUEL_AGE_NILDEPDiesel&Others5NO    0.62298    0.12371   5.036 4.76e-07 ***
FUEL_AGE_NILDEPDiesel&Others5YES   0.18147    0.19918   0.911 0.362247
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Theta: count = 0
```

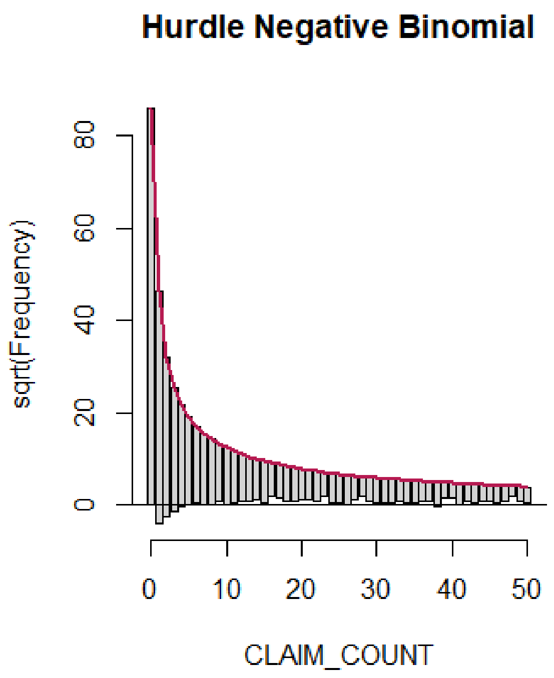
AIC comparison of Poisson and Negative Binomial with and without offset

Hrd_Pois	Hrd_Pois_With_offset	Hrd_NB	Hrd_NB_with_offset
741114.21	68799.05	74757.48	57574.12

Here the offset term is very useful as the data has no exposure.

Visualization of Goodness of Fit





Hurdle Model Poisson is showing similar not good at all fit for the rest the value except Zero. Though Negative Binomial is showing much better result than Poisson it is underestimating the part major part of Claim Count 1-5

Table Comparison of models

Pois	Negbin	Hurdle_pois	Hurdle_negbin	Zero_inf_pois	Zero_inf_negbin
59625.34	48286.51	68799.05	57574.12	741151.01	76046.80

As we can see that the AIC score is too high for dataset as there is a long tail(outlier) and dataset is excessively huge for the Zero-inflated and hurdle models' requirement. We now are trying for a sample dataset. We took 1000 sample from the dataset and try to fit normal Poisson linear model, Negative Binomial, Hurdle, Zero Inflated model.

Simulation Study

AIC

Pois	Negbin	Hurdle_pois	Hurdle_Negbin	Zero_inf_pois	zero_inf_negbin
3365.738	3090.585	34088.542	4610.077	34088.032	4651.254

Here we can see that the Simulation of size of 1000 shows that the Poisson and negative Binomial GLM models are perfect fit for the data, Zero Inflated models are nit so good fit as there is outliers. We test for outlier.

(outlier plot)

Here, we truncated the data upto the Claim count 20 and considered the data set.

Pois	Negbin	Hurdle_Pois	Hurdle_Negbin	Zero_Inf_Pois	Zero_Inf_Negbin
35610.11	35134.47	62861.89	50551.18	62861.51	50572.88

It can be seen that there is a significant difference because of the truncation. As the data is highly positively skewed, it is justified to take the truncated for further modelling.

Simulation Study

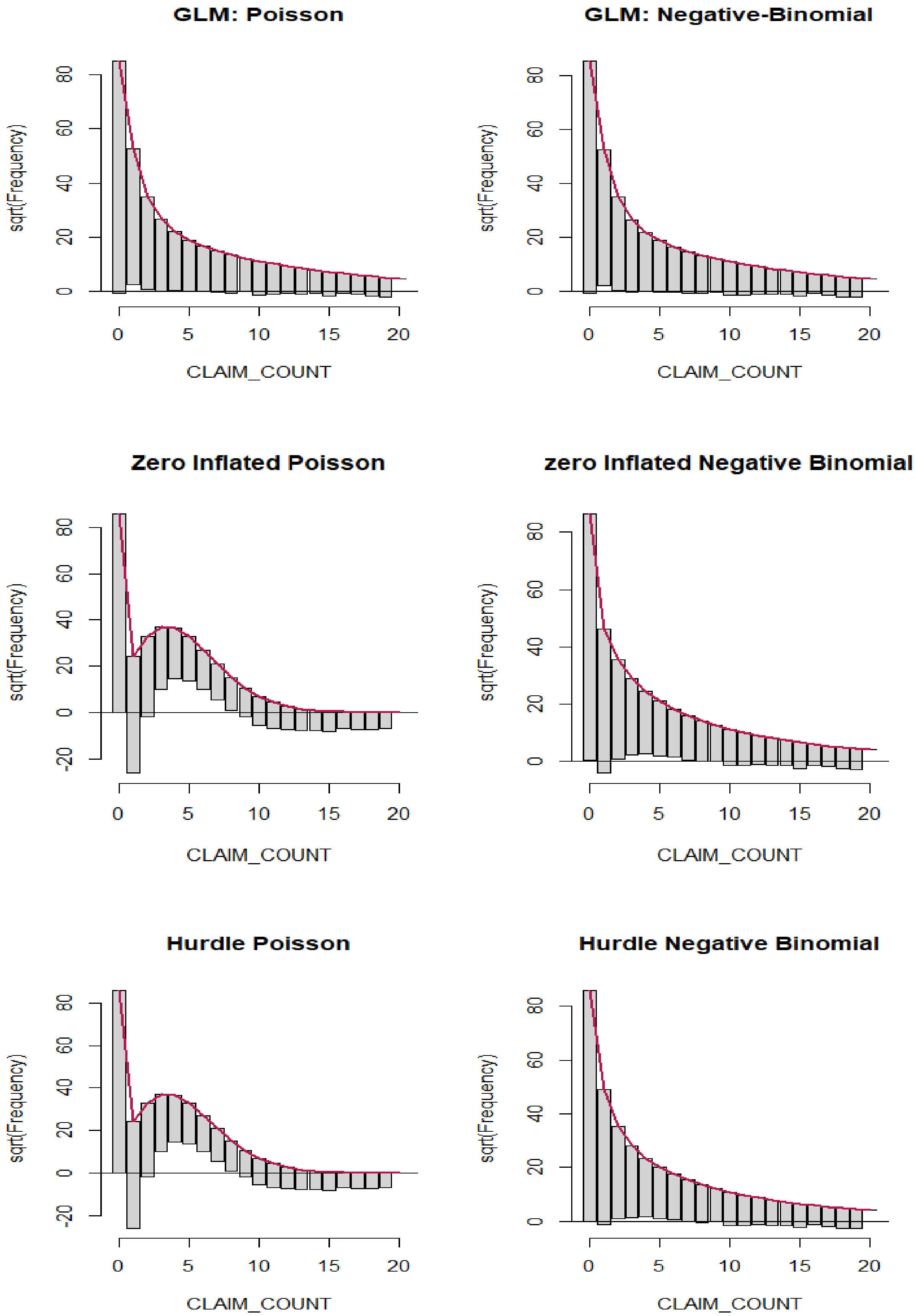
Sample	AIC_Pois	AICc_Pois	BIC_Pois	AIC_NB	AICc_NB	BIC_NB	NB>P
1000	2432	2435	2594	2420	2422	2582	100
2000	4960	4961	5145	4898	4899	5083	100
3000	7309	7310	7507	7227	7227	7425	98
4000	9817	9818	10030	9707	9707	9914	97
5000	12120	12120	12330	11970	11970	12180	100
6000	14630	14630	14850	14430	14430	14650	100
7000	17180	17180	17400	16950	16950	17180	99
8000	19720	19730	19960	19470	19470	19700	100
9000	22260	22260	22490	21940	21940	22180	100
10000	24530	24530	24770	24230	24230	24470	99

We consider the sample size of 1000 to 10000 with simulation of 100 times. Here we are comparing the goodness of fit measures AIC, AIC Corrected, BIC. BIC scores are more effected by the number of levels of the categorical variables, so it has higher values than AIC values.

We can see that the data showing Negative Binomial is better in this simulation study.

Here we are considering the truncated data to check the model goodness of fit as the outliers are less influential according to the outlier test shown above.

Visualization of Goodness of fit



We can see that the Poisson and Negative Binomial is almost similarly fitting the data in case of GLM. Though we can see some little bit difference in 0-2 Claim count data. There Negative Binomial is fitting relatively somewhat better.

The Hurdle Negative Binomial and GLM Negative binomial fit the zero counts perfectly. All of the negative binomial models show a reasonable fit, and none show a systematic pattern of under/overfitting.

If the underlying subject matter theory leads to a ZIP or a hurdle model, then that model should be applied. However, the results show that we cannot reject a GLM in favor of a ZIP or a hurdle model only because the data contain a high proportion of zeros; overdispersion, high correlation, and a covariate may well be able to explain the excessive zeros.

Validation of the model

Validation tests the predictive ability of different models by splitting the data into training and testing sets and this helps check for overfitting.

Cross-Validation

The goal of cross-**validation** is to estimate the expected level of **fit** of a **model** to a data set that is independent of the data that were used to train the **model**. It can be used to estimate any quantitative measure of **fit** that is appropriate for the data and **model**.

Poisson

MSE

3.295353	3.291882

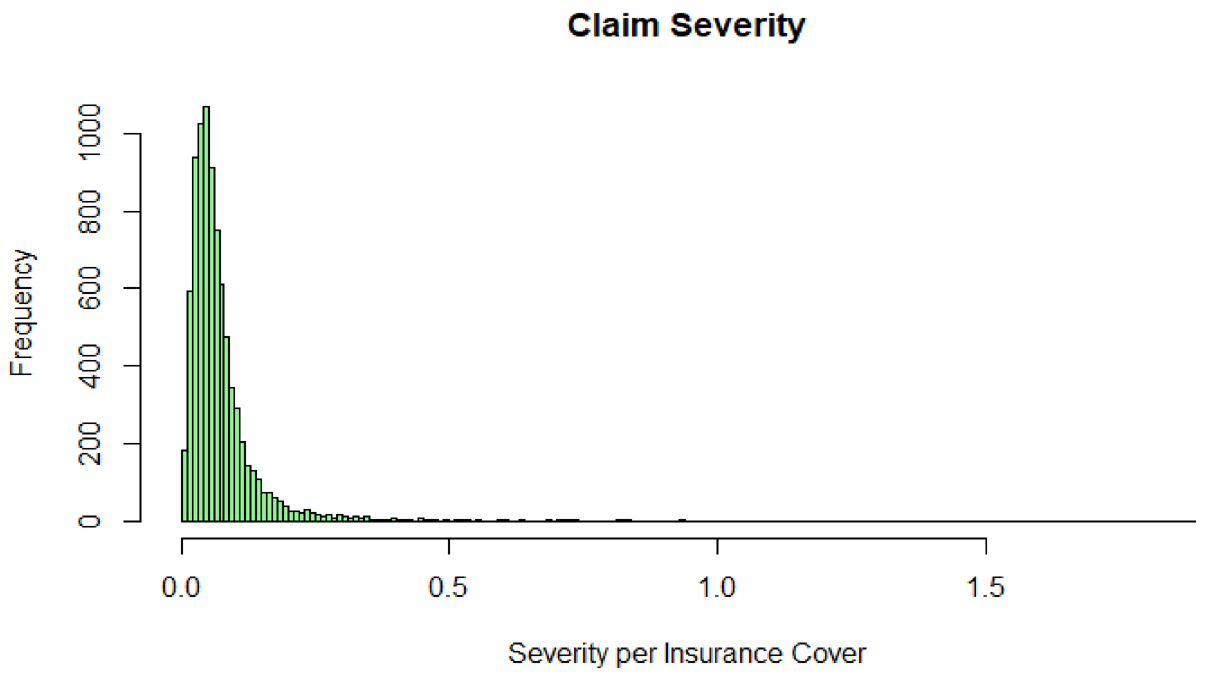
Negative Binomial

MSE

3.455130	3.451897

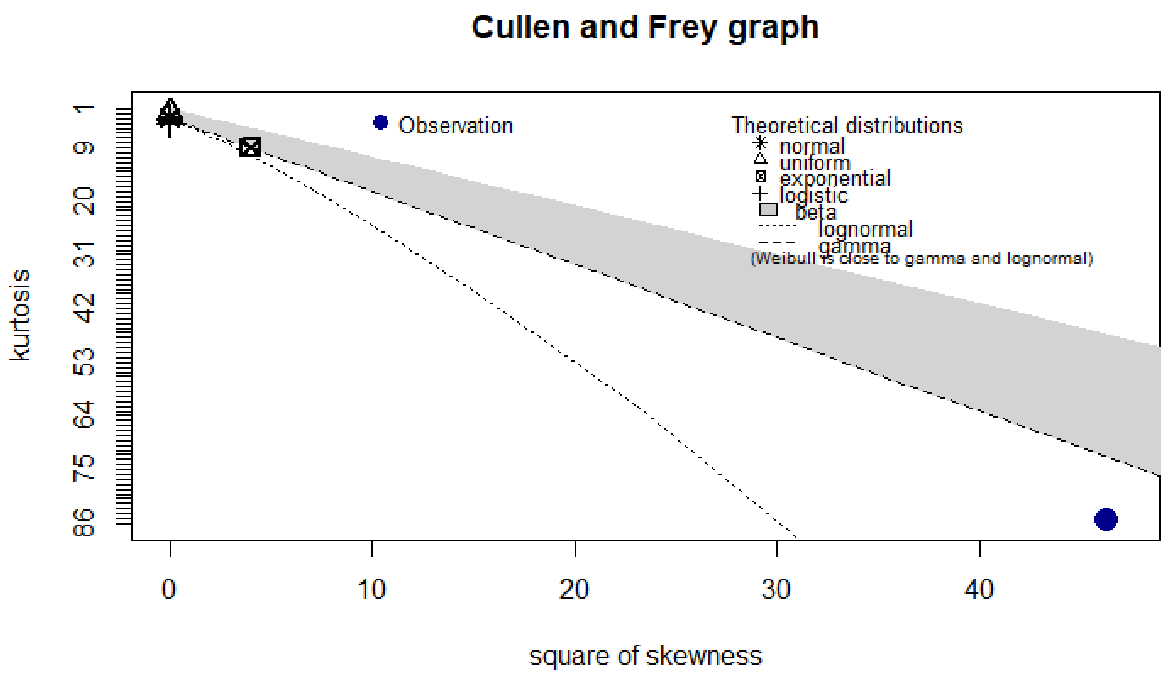
The first term in this validation is the measure of the cross validation MSE and next term is comparison of LOOCV method.

Severity



As we can see that the data is highly positive skewed. A justified explanation is because of the low claim cost is in high frequency. High frequency of the cover between 2-5 and 5-10 have been seen. So, the claim cost for those cover may not exceed the cover whereas for 50 lac cover may not generate so much claims.

Distribution



Here the distribution is closer to the Gamma distribution.

```
summary statistics
------
min:  0.000593142   max:  1.8814
median:  0.05423506
mean:  0.07331727
estimated sd:  0.0834238
estimated skewness:  6.802486
estimated kurtosis:  85.3701
```

Mean is less than variance and high positive skewness is present in data. So Gamma is perfect for the data.

Fit gamma distribution in the data

Simulation Study

```
  Gamma     Inv_Gaussian                 Gamma                    Inv Gaussian
 -3245.731  -3379.442                    -28805.485               -28833.046
```

We considered the sample of 2000 to make a basic model. Similarly,we considered the inverse gaussian instead of alternate log normal distribution.

We can see that the sample data follows have more fit for gamma distribution.

With Weights

Weights are used to make the response variable more relevant. Here, the claim counts are used as weights in the model.

```
Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                    -2.826731   0.200002 -14.133  < 2e-16 ***
CC_BAND_GROUP1001 to 1300        -0.170570   0.078249  -2.180  0.02951 *
CC_BAND_GROUP1301 to 1500        -0.296916   0.090310  -3.288  0.00105 **
CC_BAND_GROUP1501 to 2000        -0.275188   0.093036  -2.958  0.00317 **
VEHICLE_MAKEMakeB                0.210678   0.101190   2.082  0.03761 *
VEHICLE_MAKEMakeC                0.342574   0.123856   2.766  0.00579 **
VEHICLE_MAKEMakeD                0.333658   0.132690   2.515  0.01208 *
VEHICLE_MAKEMakeE                0.198752   0.087809   2.263  0.02383 *
VEHICLE_MAKEOthers              0.252207   0.086708   2.909  0.00371 **
FUEL_AGE_NILDEP1Petrol<1YES       -0.006044   0.242324  -0.025  0.98011
FUEL_AGE_NILDEP1Petrol1NO        -0.203905   0.236329  -0.863  0.38846
FUEL_AGE_NILDEP1Petrol1YES        0.147495   0.220582   0.669  0.50387
FUEL_AGE_NILDEP1Petrol2NO         0.190800   0.226630   0.842  0.40006
FUEL_AGE_NILDEP1Petrol2YES        0.105789   0.215898   0.490  0.62425
FUEL_AGE_NILDEP1Petrol3NO         0.258576   0.211906   1.220  0.22267
FUEL_AGE_NILDEP1Petrol3YES        0.287459   0.212691   1.352  0.17684
FUEL_AGE_NILDEP1Petrol4NO         0.249529   0.213287   1.170  0.24232
FUEL_AGE_NILDEP1Petrol4YES        0.199528   0.211491   0.943  0.34570
FUEL_AGE_NILDEP1Petrol5NO         0.428554   0.205534   2.085  0.03733 *
FUEL_AGE_NILDEP1Petrol5YES        0.060686   0.420131   0.144  0.88518
FUEL_AGE_NILDEPDiesel&Others<1NO  -0.237109   0.321354  -0.738  0.46079
FUEL_AGE_NILDEPDiesel&Others<1YES  0.056647   0.227417   0.249  0.80334
FUEL_AGE_NILDEPDiesel&Others1NO   -0.383624   0.277311  -1.383  0.16688
FUEL_AGE_NILDEPDiesel&Others1YES  -0.143400   0.222887  -0.643  0.52013
FUEL_AGE_NILDEPDiesel&Others2NO    0.439025   0.220132   1.994  0.04640 *
FUEL_AGE_NILDEPDiesel&Others2YES   0.126899   0.213215   0.595  0.55187
FUEL_AGE_NILDEPDiesel&Others3NO    0.383276   0.239291   1.602  0.10955
FUEL_AGE_NILDEPDiesel&Others3YES   0.193530   0.222792   0.869  0.38525
FUEL_AGE_NILDEPDiesel&Others4NO   -0.023080   0.234305  -0.099  0.92155
FUEL_AGE_NILDEPDiesel&Others4YES   0.155596   0.223870   0.695  0.48721
FUEL_AGE_NILDEPDiesel&Others5NO    0.278955   0.236851   1.178  0.23918
```

```
FUEL_AGE_NILDEPDiesel&Others5YES   1.022576   0.428663   2.386  0.01725 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Gamma family taken to be 3.163923)

    Null deviance: 1963.6  on 987  degrees of freedom
Residual deviance: 1737.5  on 956  degrees of freedom
AIC: -15051
```

Gamma_Sample	Inv_Gaussian	Gamma_data	Inv_Gaussian
-15051.13	-15223.86	-1151048.14	-1127996.80

For the full data it can be seen that the Gamma is better though Inv Gaussian is giving less AIC value

Simulation study

Sample Size	AIC_G	AICc_G	BIC_G	AIC_IG	AICc_IG	BIC_IG
1000	-1760	-1755	-1621	-15220	-15220	-15060
2000	-3489	-3487	-3327	-3574	-3572	-3411
3000	-5209	-5207	-5033	-5352	-5350	-5176
4000	-6715	-6714	-6531	-6791	-6790	-6608
5000	-8369	-8368	-8177	-8400	-8399	-8209
6000	-9770	-9770	-9573	-9833	-9832	-9636
7000	-11560	-11560	-11360	-11680	-11680	-11480
8000	-13250	-13250	-13040	-13270	-13270	-13060
9000	-14840	-14840	-14630	-15060	-15060	-14840
10000	-16240	-16240	-16020	-16560	-16560	-16350

Though Inverse Gaussian is giving lower AIC value, most of the time the selected sample didn't converge the algorithm. That is, in R we use Least square method to calculate the model estimate instead of Maximum likelihood. So, the method has some drawback due to Statistical Software used.

Validation of model

Gamma is more accurately validating the model in terms of RMSE calculated in Cross Validation.

RMSE

0.008182451 0.008177364

Chapter 7

Frequency Severity Rate Making

Intuition

Manual rating of specific risks begin with a base rate, which is then modified by appropriate relativity factors depending on characteristics of each risk.

The premium calculated is equal to Expected value of Cost of claim × Expected value of Frequency of claim

.

Coefficients:

	frequency_Estimate	Severity_Estimate	Exp(Frequency)	Exp(Severity)	Rate
(Intercept)	-2.005881	-2.5365331	0.13	0.08	0.10
CC_BAND_GROUP1001 to 1300	0.215205	-0.329549	1.24	0.72	0.96
CC_BAND_GROUP1301 to 1500	0.292802	-0.3902261	1.34	0.68	0.89
CC_BAND_GROUP1501 to 2000	0.274285	-0.4660856	1.32	0.63	0.56
VEHICLE_MAKEMakeB	-0.117838	0.1475886	0.89	1.16	0.92
VEHICLE_MAKEMakeC	-0.236242	0.2962052	0.79	1.34	1.07
VEHICLE_MAKEMakeD	-0.229165	0.071452	0.80	1.07	1.12
VEHICLE_MAKEMakeE	0.043735	0.0839514	1.04	1.09	0.98
VEHICLE_MAKEOthers	-0.106051	0.2392754	0.90	1.27	2.30
FUEL_AGE_NILDEP1Petrol<1YES	0.594626	-0.3363759	1.81	0.71	0.72
FUEL_AGE_NILDEP1Petrol1NO	0.004302	0.0156701	1.00	1.02	1.93
FUEL_AGE_NILDEP1Petrol1YES	0.642648	-0.1552366	1.90	0.86	0.88
FUEL_AGE_NILDEP1Petrol2NO	0.031544	0.0808607	1.03	1.08	2.15
FUEL_AGE_NILDEP1Petrol2YES	0.682679	-0.0007629	1.98	1.00	1.10
FUEL_AGE_NILDEP1Petrol3NO	0.091929	0.084625	1.10	1.09	2.28
FUEL_AGE_NILDEP1Petrol3YES	0.741483	0.1109149	2.10	1.12	1.20
FUEL_AGE_NILDEP1Petrol4NO	0.075008	0.1569177	1.08	1.17	2.75
FUEL_AGE_NILDEP1Petrol4YES	0.854847	0.1544955	2.35	1.17	1.02
FUEL_AGE_NILDEP1Petrol5NO	-0.136161	0.4878815	0.87	1.63	2.90
FUEL_AGE_NILDEP1Petrol5YES	0.576014	0.1122549	1.78	1.12	1.50
FUEL_AGE_NILDEPDiesel&Others<1NO	0.292826	-0.1083215	1.34	0.90	2.00
FUEL_AGE_NILDEPDiesel&Others<1YES	0.801574	-0.3362852	2.23	0.71	0.90
FUEL_AGE_NILDEPDiesel&Others1NO	0.236378	-0.0637727	1.27	0.94	2.27
FUEL_AGE_NILDEPDiesel&Others1YES	0.884777	-0.1353822	2.42	0.87	1.18
FUEL_AGE_NILDEPDiesel&Others2NO	0.304801	-0.0099711	1.36	0.99	2.47
FUEL_AGE_NILDEPDiesel&Others2YES	0.914969	-0.0032433	2.50	1.00	1.27
FUEL_AGE_NILDEPDiesel&Others3NO	0.243463	0.0199015	1.28	1.02	2.61
FUEL_AGE_NILDEPDiesel&Others3YES	0.939971	0.0761507	2.56	1.08	1.41

R Codes

```
with(GLM_data_freq, c(mean = mean(CLAIM_COUNT), var = var(CLAIM_COUNT), ratio =
var(CLAIM_COUNT) / mean(CLAIM_COUNT)))
```

```
t=factor(CLAIM_COUNT,levels=0:20)
clm.tab=table(t)
clm.tab
par(mfrow=c(1,2))
barplot(clm.tab, xlab = "Number of CLAIMS", ylab = "Frequency",col = "lightblue")

abline(v = mean(CLAIM_COUNT), col = "red", lwd = 3)
ci <- mean(CLAIM_COUNT) + c(-1, 1) * sd(CLAIM_COUNT)
lines(x = ci, y = c(-4, -4), col = "red", lwd = 3, xpd = TRUE)

barplot(clm.tab , ylab = "log(Frequency)", xlab = "NUmber of CLAIMs", col = "lightblue", log = "y")
abline(v = mean(CLAIM_COUNT), col = "red", lwd = 3)
ci <- mean(CLAIM_COUNT) + c(-1, 1) * sd(CLAIM_COUNT)
lines(x = ci, y = c(-4, -4), col = "red", lwd = 3, xpd = TRUE)
```

Generalized Linear Model

```
obj2=glm(CLAIM_COUNT~CC_BAND_GROUP+VEHICLE_MAKE+FUEL_AGE_NILDEP+offset(
log(NCB_ADJ_POLICY_COUNT)), data=tran_data[,c(-1,-17)],family=poisson(log))
```

```
library(MASS)
obj4=glm.nb(CLAIM_COUNT~CC_BAND_GROUP+VEHICLE_MAKE+FUEL_AGE_NILDEP+offs
et(log(NCB_ADJ_POLICY_COUNT)), data=tran_data[,c(-1,-17)])
```

Zero Inflated

```
library(VGAM)
 set.seed(1234)
 data1 <- rzipois(200, 3, 0)
 data2 <- rzipois(200, 3, .5)
 par(mfrow=c(1,2))
 tdata1 <- table(data1)
  barplot(tdata1, xlab = "Count", ylab = "Frequency", main = "Poisson(3)")
  tdata2 <- table(data2)
  barplot(tdata2, xlab = "Count", ylab = "Frequency",main = expression("ZI Poisson(3, " * pi * "= .5)"))

library(pscl)
zip2=zeroinfl(CLAIM_COUNT~CC_BAND_GROUP+VEHICLE_MAKE+FUEL_AGE_NILDEP,data
=tran_data,dist="poisson",link="logit", control=zeroinfl.control("BFGS"))
```

```
zinb4=zeroinfl(CLAIM_COUNT~CC_BAND_GROUP+VEHICLE_MAKE+FUEL_AGE_NILDEP,data=tran_data,dist="negbin",link="logit", control=zeroinfl.control("L-BFGS-B"))
```

Hurdle Model

```
ctrl <- hurdle.control(method = "L-BFGS-B")
ctrl$reltol <- NULL
hrd_p2=hurdle(CLAIM_COUNT~CC_BAND_GROUP+VEHICLE_MAKE+FUEL_AGE_NILDEP,data=tran_data,dist="poisson",zero.dist = "binomial",link="logit",control=ctrl)


ctrl <- hurdle.control(method = "L-BFGS-B")
ctrl$reltol <- NULL
hrd_nb4=hurdle(CLAIM_COUNT~CC_BAND_GROUP+VEHICLE_MAKE+FUEL_AGE_NILDEP,data=tran_data,dist="negbin",zero.dist = "binomial",link="logit",control=ctrl)
```

Rootogram

```
par(mfrow=c(1,2))
 countreg::rootogram(obj2, max = 20,main = "GLM: Poisson")
 countreg::rootogram(obj4, max = 20, main = "GLM: Negative-Binomial")
 rootogram(zip2, max = 20,main = "Zero Inflated Poisson")
 countreg::rootogram(zinb4, max = 20, main = "Zero Inflated Negative Binomial")
 rootogram(hrd_p2, max = 20,main = "Hurdle Poisson")
 countreg::rootogram(hrd_nb4, max = 20, main = "Hurdle Negative Binomial")
```

Test for model comparison

```
vuong(zip2,zinb4)

k=list("pois"=obj2,"negbin"=obj4,"Hurdle_pois"=hrd_p4,"Hurdle_negbin"=hrd_nb4,"zero_inf_pois"=zip2,"zero_inf_negbin"=zinb4)
#sapply(k,function(x)coef(x))
sapply(k,function(x)AIC(x))

LRstats(obj2,obj4,hrd_p4,hrd_nb4,zip2,zinb4,sortby="AIC")
```

Severity Modeling

```
obj_sev1=glm(Severity_by_IDV~CC_BAND_GROUP+VEHICLE_MAKE+FUEL_AGE_NILDEP,
family=Gamma(link=log),data=samp1[samp1$Severity>0,])

obj_sev2=glm(Severity_by_IDV~CC_BAND_GROUP+VEHICLE_MAKE+FUEL_AGE_NILDEP,
family=inverse.gaussian(link=log),data=samp1[samp1$Severity>0,])

#fitting glm of severity to whole data

obj_g=glm(Severity_by_IDV~CC_BAND_GROUP+VEHICLE_MAKE+FUEL_AGE_NILDEP,
family=Gamma(link=log),data=GLM_data_severity[GLM_data_severity$Severity>0,])

obj_ig=glm(Severity_by_IDV~CC_BAND_GROUP+VEHICLE_MAKE+FUEL_AGE_NILDEP,
family=inverse.gaussian("log"),data=GLM_data_severity[GLM_data_severity$Severity>0,])

w=list("gamma_s"=obj_sev1,"inv_g"=obj_sev2,"Gamma_data"=obj_g, "Inv.gauss"=obj_ig)
sapply(w,function(x)AIC(x))
```

AIC BIC comparison of sample size

```
R=1
n=10
Pois=data.frame()
Neg_bin=data.frame()
Gamma=data.frame()
Inverse_gaus=data.frame()
freq_model=data.frame()
sev_model=data.frame()
Count_freq=rep(0,n)
Count_sev=rep(0,n)
for (j in 1:n){

 cnt=0
 cntt=0
 pois=data.frame()
 neg=data.frame()
 gam=data.frame()
 inv.g=data.frame()

 for(i in 1:R){
 tryCatch({ ind=sample(1:nrow(tran_data),1000*j,replace = F)
  ind=sample(1:nrow(tran_data),1000*j,replace = F)
  samp_ind=tran_data[ind,]

obj2=glm(CLAIM_COUNT~CC_BAND_GROUP+VEHICLE_MAKE+FUEL_AGE_NILDEP+offset(
log(NCB_ADJ_POLICY_COUNT)), data=samp_ind[,c(-1,-17)],family=poisson(log))


obj4=glm.nb(CLAIM_COUNT~CC_BAND_GROUP+VEHICLE_MAKE+FUEL_AGE_NILDEP+offs
et(log(NCB_ADJ_POLICY_COUNT)), data=samp_ind[,c(-1,-17)])
 if(AIC(obj2)>AIC(obj4)){
  cnt=cnt+1
  }
 },error=function(e){})

  pois=rbind(pois,compareGLM(obj2,obj4)$Fit.criteria[1,3:5])
  neg=rbind(neg,compareGLM(obj2,obj4)$Fit.criteria[2,3:5])


tryCatch({obj_sev1=glm(Severity_by_IDV~CC_BAND_GROUP+VEHICLE_MAKE+FUEL_AGE_N
ILDEP, family=Gamma(link=log),data=samp_ind[which(samp_ind$Severity>0),])

  obj_sev2=glm(Severity_by_IDV~CC_BAND_GROUP+VEHICLE_MAKE+FUEL_AGE_NILDEP,
family=inverse.gaussian(link=log),data=samp_ind[samp_ind$Severity>0,])

  if(AIC(obj_sev1)<AIC(obj_sev2)){
   cntt=cntt+1
  }
  },error=function(e){})
  gam=rbind(gam,compareGLM(obj_sev1,obj_sev2)$Fit.criteria[1,3:5])
```

```
    inv.g=rbind(inv.g,compareGLM(obj_sev1,obj_sev2)$Fit.criteria[2,3:5])
}

P=data.frame(min(pois$AIC),min(pois$AICc),min(pois$BIC))
Pois=rbind(Pois,P)
N=data.frame(min(neg$AIC),min(neg$AICc),min(neg$BIC))
Neg_bin=rbind(Neg_bin,N)
G=data.frame(min(gam$AIC),min(gam$AICc),min(gam$BIC))
Gamma=rbind(Gamma,G)
I=data.frame(min(inv.g$AIC),min(inv.g$AICc),min(inv.g$BIC))
Inverse_gaus=rbind(Inverse_gaus,I)
Count_freq[j]=cnt
Count_sev[j]=cntt

}

freq_model=data.frame(Pois,Neg_bin,Count_freq)
attribut1=c("AIC_P","AICc_P","BIC_P","AIC_NB","AICc_NB","BIC_NB","NB>P")
names(freq_model)=attribut1
sev_model=data.frame(Gamma,Inverse_gaus,Count_sev)
attribut2=c("AIC_G","AICc_G","BIC_G","AIC_IG","AICc_IG","BIC_IG","G>IG")
names(sev_model)=attribut2
freq_model

sev_model
```

Validation

```
library(boot)

obj_sev1=glm(Severity_by_IDV~CC_BAND_GROUP+VEHICLE_MAKE+FUEL_AGE_NILDEP,
family=Gamma(link=log),data=tran_data[tran_data$Severity>0,])

s=cv.glm(tran_data[tran_data$Severity>0,],obj_sev1,K=10)$delta

print(s)
```

References

- A. Z., Kleiber, C., & Jackman, S. (2017). *Regression Models for Count Data in R.* Universit• at Innsbruck;Universit• at Basel;Stanford University, Statistics, Austria.
- Brisard, E. (2013-2014). Pricing of Car Insurance With Generalized Linear Models. *Vrije Universiteit Brussel*.
- Dan Tevet, M. G. (2016). *GENERALIZED LINEAR MODELS FOR INSURANCE RATING.* Casualty Actuarial Society.
- Ha, T. (2017). *Modeling the Premium in Non-Life Insurance.* university of Oslo.
- HenryPoincaré. (2017). *Generalized Linear Models for Count Data.*
- Latife Sinem Sarul, S. S. (2005). AN APPLICATION OF CLAIM FREQUENCY DATA USING ZERO INFLATED AND HURDLE MODELS IN GENERAL INSURANCE. *Journal of Business, Economics & Finance*.
- TamHa. (autumn 2017). *Modeling the Premium in Non-Life Insurance.* University of Oslo: Master's Thesis.
- Yip CHing Han, K. (2004). *Zero-Inflated Count Data Models for Claim Frequency in General Insurance.* Hong Kong.