FREQUENTIST AND BAYESIAN ZERO-INFLATED REGRESSION MODELS ON INSURANCE CLAIM FREQUENCY: A COMPARISON STUDY USING MALAYSIAN MOTOR INSURANCE DATA

Razik Ridzuan Mohd Tajuddin[1a]* and Noriszura Ismail[2a]

[a]Department of Mathematical Sciences, Faculty of Science and Technology, Universiti Kebangsaan Malaysia, Bangi, MALAYSIA. E-mail: razikridzuan@siswa.ukm.edu.my[1];ni@ukm.edu.my[2]
Corresponding Author: razikridzuan@siswa.ukm.edu.my
Received: 29th Jan 2021 Accepted: 8th Oct 2021 Published: 30th Jun 2022
DOI: https://doi.org/10.22452/mjs.vol41no2.2

ABSTRACT A no-claim event is a common scenario in insurance and the abundance of no-claim events can be described adequately using zero-inflated models. The zero-inflated Poisson (ZIP) and the zero-inflated negative binomial (ZINB) regression models from the frequentist and Bayesian approaches were compared by considering Malaysian motor insurance data as a case study. The data was categorized into three claim types and the factors considered for regression modelling were coverage type, vehicle age, vehicle cubic capacity and vehicle make. Using mean absolute deviation and mean squared prediction error values as indicators for model comparison, it was discovered that the ZIP model from both approaches fit the data better than the ZINB model. Also, both ZIP and ZINB models from the Bayesian approach fit the data better than the frequentist models. Therefore, the Bayesian ZIP model is the best model for explaining motor insurance claim frequency in Malaysia for all three claim types. According to the best regression model, the most influential factors in determining the claim frequency for each claim type are vehicle age, coverage type and vehicle make. Vehicle age and coverage type have a positive influence on the claim frequency, while vehicle make has a negative influence.

ABSTRAK Peristiwa tanpa tuntutan merupakan senario biasa yang berlaku dalam insurans dan kekerapan kejadian tanpa tuntutan yang berlebihan dapat dijelaskan dengan baik oleh model lebihan sifar. Model regresi lebihan-sifar Poisson (ZIP) dan lebihan-sifar binomial negatif (ZINB) mengikut pendekatan frekuentis dan Bayesan dibandingkan dengan mempertimbangkan data insurans motor Malaysia sebagai kajian kes. Data insurans motor dikategorikan kepada tiga jenis tuntutan dan faktor yang dipertimbangkan untuk pemodelan adalah jenis perlindungan, usia kendaraan, isian padu kenderaan dan jenis pembuatan kenderaan. Dengan menggunakan nilai min sisihan mutlak dan min ramalan ralat kuasa dua sebagai penunjuk bagi perbandingan model, didapati bahawa model ZIP dari kedua-dua pendekatan memberikan penyuaian yang lebih baik daripada model ZINB. Selain itu, kedua-dua model ZIP dan ZINB dari pendekatan Bayesan memberikan penyuaian yang lebih baik daripada model frekuentis. Oleh itu, model ZIP Bayesan dipilih sebagai model terbaik dalam menerangkan kekerapan tuntutan insurans kenderaan di Malaysia untuk ketiga-tiga jenis tuntutan. Berdasarkan model ZIP Bayesan, faktor-faktor yang paling berpengaruh dalam menentukan kekerapan tuntutan bagi setiap jenis tuntutan adalah usia kenderaan, jenis perlindungan dan jenis pembuatan kenderaan. Faktor usia kenderaan dan jenis perlindungan mempunyai kesan positif terhadap kekerapan tuntutan manakala jenis pembuatan kenderaan memberikan kesan negatif terhadap kekerapan tuntutan.

Keywords: excess zero, linear models, own damage claim, third-party bodily injury claim, third-party property damage claim.

1. INTRODUCTION

An excess of zeroes in insurance claim count data is common, which could be due to policyholders who do not file a claim. This implies that the data has a zero-inflation property that can be incorporated into the regression model as studied by Ghosh, Mukhopadhyay and Lu (2006), Liu and Powers (2012), Roohi et al. (2016) and Xie, Lin & Wei (2014). Both frequentist and Bayesian methods can be used to fit a regression model, allowing both models to be compared (Ghosh, Mukhopadhyay & Lu, 2006; Roohi et al., 2016). Roohi et al. (2016) stated that the Bayesian regression model is superior to the frequentist regression model as the former gives a smaller standard error and a smaller confidence interval. The Bayesian regression model may also result in higher coverage probability (Ghosh, Mukhopadhyay & Lu, 2006; Liu & Powers, 2012) and lower biasness (Liu & Powers, 2012). Some other models for insurance claim count data that have been employed over the years are generalized linear models (Garrido, Genest & Schulz, 2016), hurdle models (Gilenko & Miranova, 2017), zero-inflated models (Ismail & Zamani, 2013; Wagh & Kamalja, 2017) and regression models for location, shape and scale (Tzougas, Vrontos & Frangos, 2015).

Several count distributions have been used over the years, with the Poisson distribution (Garrido, Genest & Schulz, 2016; Ghosh, Mukhopadhyay & Lu, 2006; Ismail & Zamani, 2013; Wagh & Kamalja, 2017) being the most common. Negative binomial distribution (Ismail & Zamani, 2013; Tzougas, Vrontos & Frangos, 2015; Wagh & Kamlja, 2017) is used if there is overdispersion. Generalized Poisson distribution is also used by Ismail and Zamani, (2013) and Wagh and Kamalja, (2017). A generalized Poisson distribution can be used on overdispersed or underdispersed count data (Ismail & Zamani, 2013). Zero-inflated distributions such as zero-inflated Poisson (Ghosh, Mukhopadhyay & Lu, 2006; Liu & Powers, 2012; Ismail & Zamani, 2013; Rodrigues, 2003; Wagh & Kamalja, 2017, Zamani & Ismail, 2014), zero-inflated negative binomial (Ismail & Zamani, 2013; Roohi et al., 2016) and zero-inflated generalized Poisson (Ismail & Zamani, 2013; Wagh & Kamalja, 2017; Xie, Lin & Wei, 2014; Zamani & Ismail, 2014) are also widely used. The zero-inflated negative binomial and the zero-inflated generalized Poisson distributions both account for overdispersion and zero-inflation properties, but the latter also accounts for underdispersion.

Several factors are considered when modelling insurance claim count, including vehicle age, vehicle cubic capacity, vehicle make, coverage type, gender and driving experience, among others. The claim frequency decreases as the vehicle age increases (Ismail & Zamani, 2013; Wagh & Kamalja, 2017), while it increases as the vehicle cubic capacity increases (Ismail & Zamani, 2013). Zamani and Ismail (2014) discovered that policyholders with non-comprehensive coverage are more likely to file a claim for a third-party bodily injury claim. Policyholders with foreign vehicles file more claims than those with domestic vehicles (Ismail & Zamani, 2013; Zamani & Ismail, 2014). Gilenko and Miranova (2017) used vehicle class as opposed to vehicle make in their study. They discovered that policyholders with high-class vehicles are more likely to file claims. Other research has found that policyholders who have a no-claim discount advantage (Wagh & Kamalja, 2017), a lot of driving experience (Gilenko & Miranova, 2017), a high deductible (Gilenko & Miranova, 2017) and female drivers (Gilenko & Miranova, 2017; Wagh & Kamalja, 2017) are less likely to file

claims.

Prior research primarily focused on the development of regression models using either a classical or Bayesian approach, but not both, especially using Malaysian insurance data. This is because model fitting in the classical approach is usually done by maximizing log-likelihood, while model fitting in the Bayesian approach is usually done by minimizing deviance. Even though the underlying concept for estimation is similar for both approaches, the Bayesian approach uses iterations for convergence, while the classical approach uses the closed-form for the estimator directly if one can be obtained.

This paper was motivated by the need to compare zero-inflated regression models using both frequentist and Bayesian approaches for Malaysian insurance claim frequency data. For a fair comparison, mean absolute deviation (MAD) and mean squared prediction error (MPSE) are used to select the best model. The effects of certain factors can be investigated using Malaysian insurance claim frequency data as a case study to determine which model and covariate are appropriate and significant in describing the behaviour of the data.

This paper contributes to a better understanding of how to compare models with different bases of ideas using MAD and MPSE. Specifically, the classical approach estimates parameters solely based on data, while the Bayesian approach estimates the parameters based on a prior understanding of how individual factors affect the overall model. Readers will find a coding example on model fitting in the appendix, which will aid them in developing other Bayesian regression models.

The paper is organized as follows. The background information of Malaysian motor insurance and the development of classical and Bayesian zero-inflated regression models are discussed in Section 2. The results of the model fittings for various types of claims are discussed in Section 3. The final section provides a summary of the research, some concluding remarks as well as limitations and suggestions for future studies.

2. DATA AND MODELS

Malaysian motor insurance claim frequency data from 2001 to 2003 was used in this study. The covariates for modelling data were chosen based on their availability and suggestions from previous studies. Table 1 lists the covariates and their categories.

Table 1. Covariates and their categories

Covariates	Coverage type	Vehicle age	Vehicle cubic capacity	Vehicle make
1	Comprehensive	0-1 years	0-1000 cc	Local 1
2	Non-comprehensive	2-3 years	1001-1300 cc	Local 2
3		4-5 years	1301-1500 cc	Foreign 1
4		6-7 years	1501-1800 cc	Foreign 2
5		8+ years	1801+ cc	Foreign 3

The data was separated into three claim types, which are then classified into risk classes. The three types of claim data are own damage (OD), third-party bodily injury (TPBI) and third-party property damage (TPPD). Each claim type is

described by three major categorical covariates, which are vehicle age, vehicle cubic capacity and vehicle make. An additional risk factor known as coverage type was included for TPBI and TPPD claims.

The zero-inflated index proposed by Puig and Valero (2006) was used to detect the presence of excess zeroes in the data. The zero-inflated index for OD, TPBI and TPPD claim frequencies are 0.98, 0.96 and 0.92, respectively. Since the index values are positive, the three datasets have an excess of zero-valued observations, implying that the datasets have a lot of no-claim counts. The suggestion is also backed up by the dispersion index, which shows that the dispersion index for OD, TPBI and TPPD are 402, 216 and 207, respectively. It is not surprising that the dispersion indices are huge since excess zeroes in the data are some of the contributors to overdispersion (Wagh & Kamalja, 2018). The zero-inflation property for each type of claim can also be depicted visually, as shown in Figure 1, where each type of claim has a huge spike at zero counts.

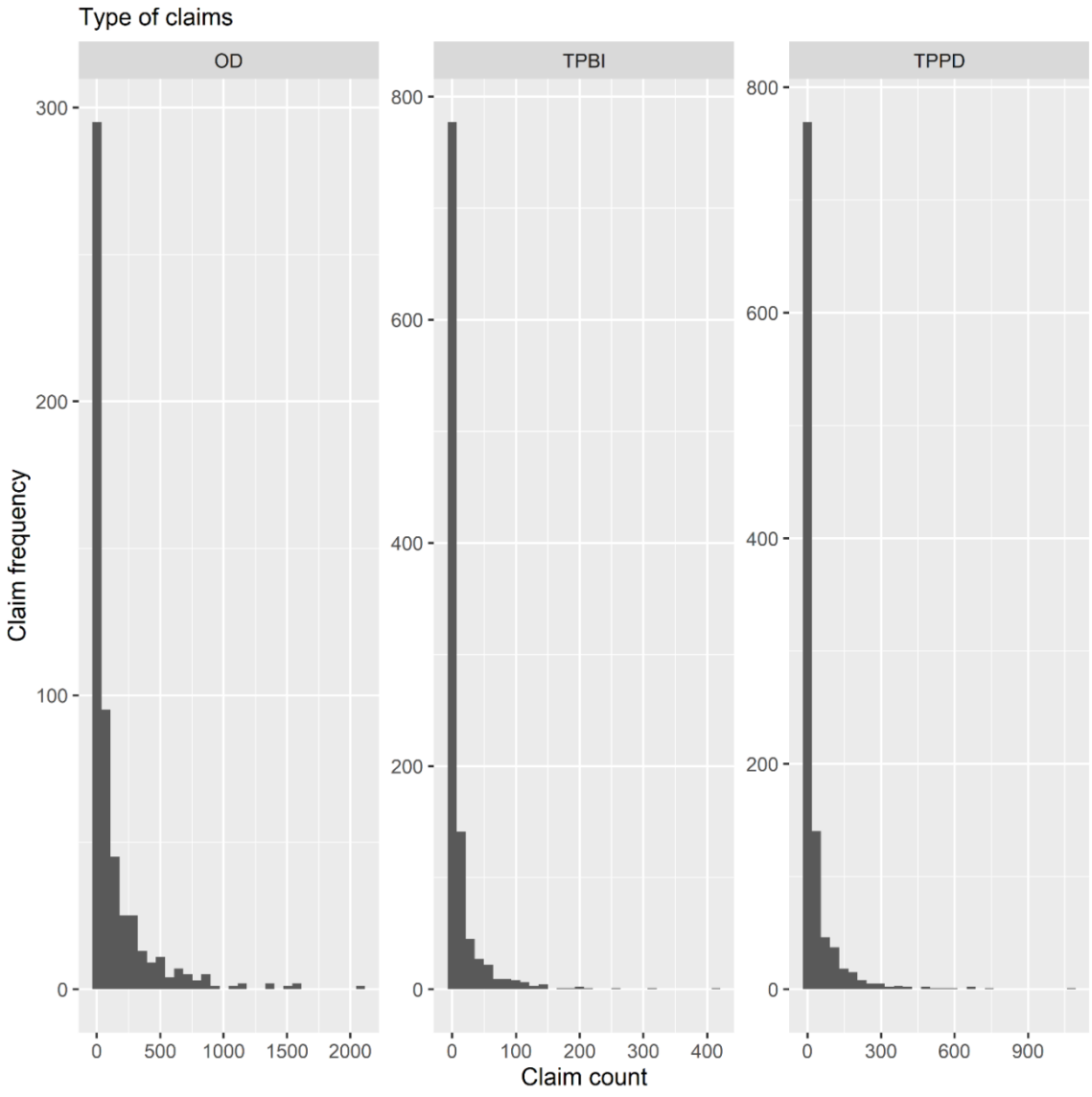

Figure 1. Claim frequency histogram for the three types of claims.

As the study aims to compare the classical and Bayesian zero-inflated

regression models, the two pioneering zero-inflated regression models, which are the zero-inflated Poisson and the zero-inflated negative binomial regression models, have been developed via classical and Bayesian approaches. When developing the regression model, it is assumed that the covariates have no direct or indirect effect on the excess in the zero counts. Hence, the covariates are only related to the mean parameter of the non-inflated distributions, where they describe the mean claim frequency for anyone who fits any combination of covariate categories.

2.1 Zero-inflated Poisson regression model

Let Y_i be a random variable that represents the number of claims in risk class i that follows a zero-inflated Poisson (ZIP) distribution with rate λ_i and proportion p_i. The mean and variance for ZIP distribution are given by $(1-p_i)\lambda_i$ and $(1-p_i)\lambda_i(1+p_i\lambda_i)$, respectively (Ismail & Zamani, 2013). The probability mass function (pmf) for the ZIP distribution for $\lambda_i, p_i > 0$ is given as

$$\Pr\left(Y_i = y_i \mid \lambda_i, p_i\right) = \begin{cases} p_i + \left(1-p_i\right)\exp\left(-\lambda_i\right) & ; y_i = 0 \\ \left(1-p_i\right)\exp\left(-\lambda_i\right)\dfrac{\lambda_i^{y_i}}{y_i!} & ; y_i > 0 \end{cases}. \quad (1)$$

The frequentist regression approach can be implemented by applying appropriate link functions. Logit link with intercept was used for proportion parameter p_i, since there is no prior information that any of the covariates considered have a direct or indirect relationship to the logit link model. The cost of including all variables in the logit link model is higher than the cost of excluding those variables. For rate parameter λ_i, log link was used. The link functions for p_i and λ_i are

$$\text{logit}\left(p_i\right) = \ln\left(\frac{p_i}{1-p_i}\right) = \tau \quad (2)$$

and

$$\ln\left(\lambda_i\right) = \ln\left(e_i\right) + \mathbf{X}_i^T\boldsymbol{\beta}, \quad (3)$$

respectively, where τ is the Bernoulli intercept, $\boldsymbol{X}$ is the vector of covariates, $\boldsymbol{\beta}$ is the vector of coefficients for covariates and e_i is the exposure data. The parameters for the link functions in (2) and (3) can be rewritten in another form as $p_i = [1+\exp(-\tau)]^{-1}$ and $\lambda_i = e_i\exp(\boldsymbol{X}_i^T\boldsymbol{\beta})$, respectively. The likelihood function for the ZIP model is

$$L\left(\mathbf{y} \mid \boldsymbol{\lambda}, \mathbf{p}\right) \propto \exp\left\{\sum_{i=1}^{k}\ln\left[p_i + \left(1-p_i\right)\exp\left(-\lambda_i\right)\right] + \sum_{i=k+1}^{n}\ln\left(1-p_i\right) - \lambda_i - y_i\ln\left(\lambda_i\right) - \ln\left(y_i!\right)\right\}, \quad (4)$$

where $i = 1,2,\ldots,k$ is the zero-valued observations and $i = k+1, k+2, \ldots, n$ is the non-zero observations. Parameters p_i and λ_i are substituted into the log of the likelihood function shown in (4), which yields the following

$$l(\mathbf{y}\mid\tau,\boldsymbol{\beta}) = \ln L(\mathbf{y}\mid\boldsymbol{\lambda},\mathbf{p}) \propto \sum_{i=1}^{k}\ln\left\{\left[1+\exp(-\tau)\right]^{-1}+\left[1+\exp(\tau)\right]^{-1}\exp\left[-e_i\exp\left(\mathbf{X}_i^T\boldsymbol{\beta}\right)\right]\right\} + \sum_{i=k+1}^{n}\left\{\ln\left[1+\exp(\tau)\right]^{-1}-e_i\exp\left(\mathbf{X}_i^T\boldsymbol{\beta}\right)-y_i\left[\ln(e_i)+\mathbf{X}_i^T\boldsymbol{\beta}\right]-\ln(y_i!)\right\} \quad (5)$$

2.2 Zero-inflated negative binomial regression model

Let Y_i be a random variable that represents the number of claims in risk class i which follows zero-inflated negative binomial (ZINB) distribution with rate λ_i, proportion p_i and dispersion r. The reparametrized pmf for ZINB distribution is

$$\Pr(Y_i=y_i\mid\lambda_i,p_i)=\begin{cases} p_i+(1-p_i)\left(\dfrac{1}{1+r\lambda_i}\right)^{1/r} & ;\ y_i=0 \\ (1-p_i)\dfrac{\Gamma(y_i+1/r)}{\Gamma(y_i+1)\Gamma(1/r)}\left(\dfrac{1}{1+r\lambda_i}\right)^{1/r}\left(\dfrac{r\lambda_i}{1+r\lambda_i}\right)^{y_i} & ;\ y_i>0 \end{cases} \quad (6)$$

The mean for reparametrized ZINB distribution is $(1-p_i)\lambda_i$, while its variance is $(1-p_i)\lambda_i(1+r\lambda_i+p_i\lambda_i)$ (Ismail & Zamani, 2013). The likelihood function for re-parameterized ZINB given in (6) is

$$L(\mathbf{y}\mid\boldsymbol{\lambda},\mathbf{p}) \propto \exp\left\{\sum_{i=1}^{k}\ln\left[p_i+(1-p_i)(1+r\lambda_i)^{-1/r}\right]\right\} \times \exp\left\{\sum_{i=k+1}^{n}\begin{bmatrix}\ln(1-p_i)+\ln\Gamma(y_i+1/r)-\ln\Gamma(y_i+1)\\ -\ln\Gamma(1/r)-(y_i+1/r)\ln(1+r\lambda_i)+y_i\ln(r\lambda_i)\end{bmatrix}\right\}, \quad (7)$$

where $i=1,2,\ldots,k$ is the zero-valued observations and $i=k+1,k+2,\ldots,n$ is the non-zero observations. Using the link functions in (2) and (3), parameters p_i and λ_i are substituted into the log of the likelihood function in (7) which yields

$$l(\mathbf{y}\mid\tau,\boldsymbol{\beta},r) = \ln L(\mathbf{y}\mid\boldsymbol{\lambda},\mathbf{p}) \propto \sum_{i=1}^{k}\ln\left\{\left[1+\exp(-\tau)\right]^{-1}+\left[1+\exp(\tau)\right]^{-1}\left[1+re_i\exp\left(\mathbf{X}_i^T\boldsymbol{\beta}\right)\right]^{-1/r}\right\} + \sum_{i=k+1}^{n}\begin{Bmatrix}\ln\left[1+\exp(\tau)\right]^{-1}+\ln\Gamma(y_i+1/r)-\ln\Gamma(y_i+1)-\ln\Gamma(1/r)\\ -(y_i+1/r)\ln\left[1+re_i\exp\left(\mathbf{X}_i^T\boldsymbol{\beta}\right)\right]+y_i\ln\left[re_i\exp\left(\mathbf{X}_i^T\boldsymbol{\beta}\right)\right]\end{Bmatrix} \quad (8)$$

The estimated coefficients were obtained using package *pscl* (Jackman et al., 2017) of R programming language. The ZIP regression model is fitted by default, but if the *distr* command in R is changed to *negbin*, the ZINB regression model is fitted instead. The package estimates the parameters by maximizing the log-likelihood function.

2.3 Bayesian approach

Prior distributions with minimal information for each covariate were considered since only the direction of each covariate's effect, i.e., positive or negative, on the mean of the claim frequency is known. Non-informative priors were used, where the coefficients, τ and $\boldsymbol{\beta}$, follow a normal distribution with a mean of 0 and a large variance. Since some of the factors may have a positive or negative effect on the mean of the claim frequency, it is fair to use the normal distribution with support from the real number line. An extra non-informative prior for dispersion r is required for the ZINB model, with the mean set to 1 and its variance set to 1000. Given that the dispersion is positive, it is reasonable to select a prior distribution for the dispersion parameter that has support on the positive real number line, such as the gamma distribution. Therefore, the non-informative priors are $\tau \sim N(0,1000)$, $\beta_j \sim N(0,1000)$ and $r \sim \Gamma(0.001,0.001)$ for $j = 1,2,\ldots,12$. The probability density function (pdf) for r, τ and $\boldsymbol{\beta}$ are

$$\Pr(r) \propto r^{-0.999} \exp(-0.001r), \tag{9}$$

$$\Pr(\tau) \propto \exp(-0.0005\tau^2) \tag{10}$$

and

$$\Pr(\boldsymbol{\beta}) \propto \exp(-0.0005\boldsymbol{\beta}^T\boldsymbol{\beta}), \tag{11}$$

respectively. The preceding pdfs have been simplified using a proportional sign that absorbs all of the proportional constants in each distribution. The joint priors for ZIP and ZINB are

$$\Pr(\tau,\boldsymbol{\beta}) \propto \exp(-0.0005\tau^2 - 0.0005\boldsymbol{\beta}^T\boldsymbol{\beta}) \tag{12}$$

and

$$\Pr(r,\tau,\boldsymbol{\beta}) \propto r^{-0.999} \exp(-0.001r - 0.0005\tau^2 - 0.0005\boldsymbol{\beta}^T\boldsymbol{\beta}), \tag{13}$$

respectively. The joint prior for the ZIP model consists of the product of prior distributions for parameters τ and $\boldsymbol{\beta}$, while the joint prior for the ZINB model consists of the product of prior distributions for parameters r, τ and $\boldsymbol{\beta}$. Each parameter is assumed to be independent of one another. The posterior distribution is generally defined as the product of the likelihood function and its corresponding prior distributions. The joint posterior for ZIP is given as

$$\Pr(\tau,\boldsymbol{\beta}\mid\mathbf{y}) \propto \exp\left\{\begin{array}{c}\sum_{i=1}^{k}\ln\left\{\left[1+\exp(-\tau)\right]^{-1}+\left[1+\exp(\tau)\right]^{-1}\exp\left[-e_i\exp\left(\mathbf{X}_i^T\boldsymbol{\beta}\right)\right]\right\} \\ +\sum_{i=k+1}^{n}\left\{\ln\left[1+\exp(\tau)\right]^{-1}-e_i\exp\left(\mathbf{X}_i^T\boldsymbol{\beta}\right)-y_i\left[\ln(e_i)+\mathbf{X}_i^T\boldsymbol{\beta}\right]-\ln(y_i!)\right\} \\ -0.0005\tau^2-0.0005\boldsymbol{\beta}^T\boldsymbol{\beta}\end{array}\right\}, \tag{14}$$

while the joint posterior for ZINB is given as

$$\Pr(\tau,\boldsymbol{\beta},r\,|\,\mathbf{y}) \propto \exp\left\{\begin{array}{c}\displaystyle\sum_{i=1}^{k}\ln\left\{\left[1+\exp(-\tau)\right]^{-1}+\left[1+\exp(\tau)\right]^{-1}\left[1+re_i\exp\left(\mathbf{X}_i^T\boldsymbol{\beta}\right)\right]^{-1/r}\right\} \\ +\displaystyle\sum_{i=k+1}^{n}\left\{\begin{array}{c}\ln\left[1+\exp(\tau)\right]^{-1}+\ln\Gamma\left(y_i+1/r\right)-\ln\Gamma\left(y_i+1\right)-\ln\Gamma\left(1/r\right) \\ -\left(y_i+1/r\right)\ln\left[1+re_i\exp\left(\mathbf{X}_i^T\boldsymbol{\beta}\right)\right]+y_i\ln\left[re_i\exp\left(\mathbf{X}_i^T\boldsymbol{\beta}\right)\right]\end{array}\right\} \\ -0.0005\tau^2-0.0005\boldsymbol{\beta}^T\boldsymbol{\beta}-0.001r-0.999\ln(r)\end{array}\right\}. \quad (15)$$

The Markov chain Monte Carlo simulation technique was used to estimate the coefficients for covariates in the Bayesian approach. Package R2jags developed by Su and Yajima (2015) was used to estimate parameters in a Bayesian context. The coefficients were estimated using this package by minimizing the deviance. Appendix A contains a reference algorithm for estimating parameters of the ZIP regression model for OD claims using the Bayesian approach. The algorithms can be modified for the Bayesian ZINB regression model. A total of 100,000 iterations were obtained for the estimates, with the first 50,000 iterations discarded. For the remaining 50,000 iterations, every alternate iteration was kept for estimating the coefficients of the parameters.

3. RESULTS AND DISCUSSIONS

The results of model fittings based on ZIP and ZINB models using both classical and Bayesian approaches for OD, TPBI and TPPD claim count data are summarized in Table 2, Table 3 and Table 4, respectively. Both frequentist and Bayesian regression models were compared using MAD and MPSE with their respective formulae: $MAD = 1/n\sum_{i=1}^{n}|y_i-\hat{y}_i|$ and $MPSE = 1/n\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$, where $\hat{y}_i$ is the predicted claim count for policyholders in the i^{th} risk class. The model with the smallest MAD and MPSE was selected as the best model for describing the insurance claim frequency.

According to Table 2, the regression estimates for the same distribution-based models are similar only for estimates that are significant at a 5% significance level. The Bayesian zero-inflated Poisson model is the best since it gives the smallest MAD and MPSE. The coefficient of vehicles aged 2 to 3 years for OD claims indicates that policyholders who own a vehicle that is 2 to 3 years old have the highest tendency to file claims. On the other hand, those with Local 2 vehicle make have the lowest tendency to file claims. The OD claim frequency decreases as the vehicle age increases. The estimated proportion of no-claim is 0% under the Bayesian ZIP model, indicating that the model is overfitting the data. This may also suggest that the excess zeroes in the data are sufficiently explained when the covariates are included.

Table 2. Parameter estimates for frequentist and Bayesian regression (OD data)

Covariates	ZIP		ZINB	
	Frequentist	Bayesian	Frequentist	Bayesian
Intercept	-3.014*	-2.701*	-3.263*	-2.813*
2–3 years	0.549*	0.548*	0.666*	0.660*
4–5 years	0.531*	0.530*	0.588*	0.586*
6–7 years	0.453*	0.451*	0.457*	0.447*
8+ years	0.248*	0.238*	0.357*	0.341*
1001–1300 cc	-0.123*	-0.419*	0.093	-0.283*
1301–1500 cc	0.061*	-0.249*	0.433*	0.025
1501–1800 cc	0.318*	0.519	0.530*	-3.838
1801+ cc	0.380*	-0.521	0.641*	3.939
Local 2	-0.296*	-0.570*	-0.211*	-0.404*
Foreign 1	-0.253*	-0.229*	-0.324*	-0.348*
Foreign 2	0.148*	0.174*	0.279*	0.269*
Foreign 3	-0.059*	0.008	-0.234*	-0.116
$\boldsymbol{\tau}$	-13.250	-28.906*	-13.710	-29.600
$\boldsymbol{r}$	-	-	7.250	6.030
Log-likelihood	-3983.70	-4108.96	-2183.73	-2857.21
MAD	35.16	25.78	51.44	33.24
MPSE	6095.04	2908.77	14 850.79	5905.65

* significant at 5%

According to Table 3, the regression estimates for the same distribution-based models are similar only for estimates that are significant at a 5% significance level. The Bayesian zero-inflated Poisson model is the best since it gives the smallest MAD and MPSE. The policyholders for TPBI claims who owned a vehicle that is more than 8 years old have the highest tendency to file claims. On the other hand, those with Foreign 1 vehicle make have the lowest tendency to file claims. The TPBI claim frequency increases as the vehicle age increases. The estimated proportion of no-claim approximates 3.6% under the Bayesian ZIP model.

Table 3. Parameter estimates for frequentist and Bayesian regression (TPBI data)

Covariates	ZIP		ZINB	
	Frequentist	Bayesian	Frequentist	Bayesian
Intercept	-5.947*	-5.916*	-5.863*	-5.744*
Non-comprehensive	0.397*	0.395*	1.111*	1.117*
2–3 years	1.301*	1.300*	1.401*	1.409*
4–5 years	1.498*	1.497*	1.576*	1.584*
6–7 years	1.506*	1.505*	1.178*	1.186*
8+ years	1.510*	1.508*	1.082*	1.081*
1001–1300 cc	-0.051	-0.079*	0.084	-0.014
1301–1500 cc	0.050	0.021	0.342*	0.229*
1501–1800 cc	0.180*	1.006	0.564*	-8.312
1801+ cc	0.040	-0.856	0.183	8.760
Local 2	0.013	-0.013	-0.418*	-0.488*
Foreign 1	-0.229*	-0.227*	-0.119	-0.122
Foreign 2	0.070	0.073*	0.327*	0.319*
Foreign 3	-0.213*	-0.203*	-0.449*	-0.407*
τ	-3.253*	-3.293*	-12.940	-30.185*
r	-	-	1.178	1.151
Log-likelihood	-4156.35	-4188.94	-2399.59	-2415.74
MAD	5.28	5.26	9.40	9.09
MPSE	160.27	159.33	600.25	555.34

* significant at 5%

According to Table 4, the regression estimates for the same distribution-based models are similar only for estimates which are significant at a 5% significance level. The Bayesian zero-inflated Poisson model is the best since it gives the smallest MAD and MPSE. The policyholders for TPPD claims who own a vehicle that is 6 to 7 years old have the highest tendency to file claims. On the other hand, those with Local 2 vehicle make have the lowest tendency to file claims. The estimated proportion of no-claim approximates 0.5% under the Bayesian ZIP model.

Table 4. Parameter estimates for frequentist and Bayesian regression (TPPD data)

Covariates	ZIP		ZINB	
	Frequentist	Bayesian	Frequentist	Bayesian
Intercept	-4.446*	-4.111*	-4.507*	-4.166*
Non-comprehensive	0.463*	0.448*	1.169*	1.201*
2-3 years	0.963*	0.958*	1.095*	1.095*
4-5 years	1.018*	1.015*	1.070*	1.066*
6-7 years	1.045*	1.041*	0.724*	0.721*
8+ years	0.781*	0.764*	0.454*	0.435*
1001-1300 cc	-0.024	-0.339*	0.160*	-0.126
1301-1500 cc	0.129*	-0.194*	0.601	0.271*
1501-1800 cc	0.359*	0.705	0.761*	9.630
1801+ cc	0.409*	-0.676	0.521*	-9.210
Local 2	-0.338*	-0.625*	-0.615*	-0.778*
Foreign 1	-0.378*	-0.359*	-0.237*	-0.253*
Foreign 2	-0.276*	-0.253*	-0.134	-0.147
Foreign 3	-0.523*	-0.440*	-0.504*	-0.395*
τ	-5.182*	-5.306*	-12.990	-30.270*
r	-	-	1.366	1.320
Log-likelihood	-7251.97	-7325.05	-3151.29	-3208.20
MAD	13.35	11.07	25.57	16.53
MPSE	1091.81	716.23	4929.45	1878.08

* significant at 5%

4. CONCLUSIONS

The purpose of this study is to compare frequentist regression models with Bayesian regression models using the Malaysian motor insurance claim count data. Two regression models, namely zero-inflated Poisson and zero-inflated negative binomial models were developed using classical and Bayesian approaches. The existence of zero inflation in the data has been identified. This study focused on three types of motor insurance claims in Malaysia from 2001 to 2003, which are OD, TPBI and TPPD claims. Four factors were considered in the development of the regression models, namely coverage type, vehicle age, vehicle cubic capacity and vehicle make. Since there is no prior information on how these factors affect the frequency of count, non-informative priors for the coefficients of the covariates were chosen.

The model fittings showed that the estimated regression parameters for the

same model are similar if the estimates are significant at a 5% significance level. Based on the MAD and MPSE, zero-inflated Poisson models outperform zero-inflated binomial negative models for both frequentist and Bayesian approaches. However, when the frequentist and Bayesian approaches were compared, models from the Bayesian approach outperform those from the frequentist approach. The frequentist models resulted in larger log-likelihood compared to Bayesian models since the latter models include non-informative priors. It is believed that the log-likelihood for Bayesian models will improve if the right informative prior is applied.

For each type of claim, the regression models indicate that vehicle age, coverage type and vehicle make are significant in determining the claim frequency. While vehicle age and coverage type have a positive effect on the claim frequency, vehicle make has a negative effect. In other words, coverage type and vehicle age increase the mean of the claim frequency, while vehicle make decreases the mean of the claim frequency. The vehicle cubic capacity does not exhibit a consistent and significant effect on the claim frequency as shown in Tables 2 to 4.

It is important to note here that the study can be further improved by considering the characteristics of policyholders, such as gender, age, prior claim experience and so on. Such data are unavailable, but it is plausible that these characteristics can serve as important indicators in obtaining better estimates of the claim frequency. Other distributions, especially zero-inflated generalized Poisson, should be considered as a potential candidate for developing regression models and fittings. Before Bayesian modelling can be executed, the selection of the proper prior distributions for prior parameters of the zero-inflated generalized Poisson regression models must be thoroughly investigated. With prior knowledge from the experts or via the empirical Bayesian method, the use of better and more informative prior distributions can be considered.

5. ACKNOWLEDGEMENT

The authors gratefully acknowledge the financial support received in the form of research grants (FRGS/1/2019/STG06/UKM/01/5) from the Ministry of Education, Malaysia and (GUP-2019-031) from Universiti Kebangsaan Malaysia.

6. REFERENCES

Garrido, J., Genest, C. & Schulz, J. (2016). Generalized linear models for dependentfrequency and severity of insurance claims, *Insurance: Mathematics and Economics* 70: 205-215.

Ghosh, S.K., Mukhopadhyay, P. & Lu, J.C. (2006). Bayesian analysis of zero-inflatedregression models, *Journal of Statistical Planning and Inference* 136: 1360-1375.

Gilenko, E.V. & Miranova, E.A. (2017). Modern claim frequency and claim severity models: An application to the Russian motor own damage insurance market, *Cogent Economics & Finance* 5: 1311097.

Ismail, N. & Zamani, H. (2013). Estimation of claim count data using negative binomial, generalized Poisson, zero-inflated negative binomial and zero-inflated generalized Poisson regression models, *Casualty Actuarial Society E-Forum, Spring*.

Jackman, S., Tahk, A., Zeileis, A., Maimone, C., Fearon, J. & Meers, Z. (2017). Package 'pscl'.

Liu, H. & Powers, D.A. (2012). Bayesian inference for zero-inflated Poisson regression models, *Journal of Statistics: Advances in Theory and Applications* 7: 155-188.

Puig, P. & Valero, J. (2006). Count data distributions: Some characterizations with applications, *Journal of the American Statistical Association* 101: 332-340.

Rodrigues, J. (2003). Bayesian analysis of zero-inflated distributions, *Communication in Statistics – Theory and Methods* 32: 281-289.

Roohi, S., Baneshi, M.R., Norrozi, A. Hajebi, A. & Bahrampour, A. (2016). Comparing Bayesian regression and classic zero-inflated negative binomial on size estimation of people who use alcohol, *Journal of Biostatistics and Epidemiology* 2: 173-179.

Su, Y.S. & Yajima, M. (2015). Package 'R2jags'.

Tzougas, G., Vrontos, S.D. & Frangos, N.E. 2015. Risk classification for claim counts and losses using regression models for location, scale and shape, *Variance* 9: 140-157.

Wagh, Y.S. & Kamalja, K.K. (2017). Modeling auto insurance claims in Singapore, *Sri Lankan Journal of Applied Statistics* 18: 105-118.

Wagh, Y.S. & Kamalja, K.K. (2018). Zero-inflated models and estimation in zero-inflated Poisson distribution, *Communications in Statistics – Simulation and Computation* 47: 2248-2265.

Xie, F.C., Lin, J.G. & Wei, B.C. (2014). Bayesian zero-inflated generalized Poisson regression model: estimation and case influence diagnostic, *Journal of Applied Statistic*s 41: 1382-1392.

Zamani, H. & Ismail, N. (2014). Functional form for the zero-inflated generalized Poisson regression model, *Communications in Statistics – Theory and Methods* 43: 515-529.

APPENDIX A

```
#install library R2jags
library(R2jags)

#Read csv data
data            <- read.csv(file.choose(), header = T)
attach(data)

#Assign variables to data
X               <- cbind(1, data[,c(1:16)])
y               <- Odcount
E               <- exposure
n               <- length(y)
zip.data        <- list("y","E","n","X")

#Developing ZIP regression model
modelText       <- "
model{

        #Likelihood function
        for(i in 1:n){
                y[i] ~ dpois(mu[i])
                mu[i]   <- (1-u[i])*lambda[i] + 0.00001*u[i]
                log(lambda[i]) <- log(E[i]) + inprod(X[i,],beta[])

                #zero-inflation
                u[i]      ~ dbern(p[i])
                logit(p[i])       <- inprod(X[i,],alpha[])
        }

        #prior distribution
        for(j in 1:17){
                beta[j] ~ dnorm(0,0.001)
                alpha[j]          ~ dnorm(0,0.001)
        }
}"

writeLines(modelText,"ZIP.txt")

#Assigning how the estimated parameters will be displayed
zip.params      <- c(paste("beta[",i = 1:17,"]",sep = ""), paste("alpha[",i = 1:17,"]",sep= ""))

#Providing initial values of the parameters
zip.inits       <- function(){list("beta" = rep(0.1,17), "alpha" = rep(0,17))}

#Model fitting with fixed set.seed() to duplicate results (if necessary)
set.seed(10)
zip.fit  <- jags(data = zip.data, inits = zip.inits, parameters.to.save = zip.params, n.chains =
        3, n.iter = 100000, n.burnin = 50000, n.thin = 2, model.file = "ZIP.txt")
```