risks

MDPI

Article

The Applications of Generalized Poisson Regression Models to Insurance Claim Data

Pouya Faroughi [1,*], Shu Li [2] and Jiandong Ren [2]

[1] School of Mathematical and Computational Sciences, University of Prince Edward Island, Charlottetown, PE C1A 4P3, Canada
[2] Department of Statistical and Actuarial Sciences, Western University, London, ON N6A 5B7, Canada; shu.li@uwo.ca (S.L.); jren6@uwo.ca (J.R.)
* Correspondence: pfaroughi@upei.ca

Abstract: Predictive modeling has been widely used for insurance rate making. In this paper, we focus on insurance claim count data and address their common issues with more flexible modeling techniques. In particular, we study the zero-inflated and hurdle-generalized Poisson and negative binomial distributions in a functional form for modeling insurance claim count data. It is shown that these models are useful in addressing the problem of excess zeros and over-dispersion of the claim count variable. In addition, we show that including the exposure as a covariate in both the zero and the count part of the model is an effective approach to incorporating exposure information in zero-inflated and hurdle models. We illustrate the effectiveness and versatility of the introduced models using three real datasets. The results suggest their promising applications in insurance risk classification and beyond.

Keywords: risk classification; count data; over-dispersion; hurdle-generalized Poisson regression; hurdle negative binomial regression; exposure; shrinkage

Citation: Faroughi, Pouya, Shu Li, and Jiandong Ren. 2023. The Applications of Generalized Poisson Regression Models to Insurance Claim Data. *Risks* 11: 213. https://doi.org/10.3390/risks11120213

Academic Editor: Mogens Steffensen

Received: 7 November 2023
Revised: 26 November 2023
Accepted: 4 December 2023
Published: 7 December 2023

1. Introduction

In a priori risk classification, actuaries group risks with similar risk characteristics in order to set insurance premiums. Accurate risk classification is extremely important for maintaining a financially sound and equitable system, assuring the availability of needed insurance coverage to the public.

The individual risk characteristics used in risk classification are called rating variables. For example, in automobile insurance, commonly used rating variables include geography, driver characteristics such as age, gender, and marital status, and vehicle characteristics such as the make and value of the vehicle insured.

Risk classification systems are generally based, whenever possible, on statistical analysis. Naturally, statistical methods such as generalized linear models and generalized additive models provide useful tools. Numerous books and papers discuss the application of statistical methods in insurance rate making, see, e.g., Renshaw (1994), Denuit et al. (2007), Frees (2009), Frees et al. (2014), and the references therein.

This paper studies claim frequency modeling. It is well known that the Poisson regression model is not always suitable because real-world claim frequency data usually exhibit over-dispersion. Alternative models have been proposed in the literature. Notably, negative binomial regression models were discussed by Dionne and Vanasse (1989), Frees and Valdez (2008), and Wüthrich and Merz (2008). Inverse Gaussian models were studied by Dean et al. (1989) and Wang et al. (2023). Consul (1993) compared the generalized Poisson (GP) distribution with several well-known distributions and concluded that the GP distribution is a plausible model for claim frequency data.

Insurance claim data usually have an excessive number of zeros. Zero-inflated models, studied by Lambert (1992), have been used to deal with such problems in the literature.

For example, Yip and Yau (2005) applied several parametric zero-inflated count distributions, including zero-inflated Poisson (ZIP), zero-inflated generalize negative binomial, zero-inflated generalized Poisson, and zero-inflated double Poisson distributions, to accommodate the excess zeros in insurance claim count data. Famoye and Singh (2006) applied the zero-inflated generalized Poisson regression model to fit a domestic violence dataset. Czado et al. (2007) extended the zero-inflated generalized Poisson regression model by including explanatory regression parameters in both the zero-inflation and the dispersion parameters and applied the extended model to patent outsourcing rate data.

The hurdle model, which was introduced by Cragg (1971) and later refined by Mullahy (1986), can also be applied to model data with an excessive number of zeros. For instance, Saffari et al. (2013) and Zuo et al. (2021) studied the hurdle-generalized Poisson distribution, whereas Bhaktha (2018) employed the hurdle negative binomial approach. Additionally, using an insurance claim number dataset, Boucher et al. (2007) compared various zero-inflated and hurdle models.

Another issue with insurance claim datasets lies in the fact that different observations may have different risk exposures, but only the total number of claims for all exposures is recorded. For example, some policyholders stay longer in the policy than others. An "offset" term is often utilized to account for the varying exposure scale. In the case of the log link function, this is equivalent to including the log of exposure as an explanatory variable with a fixed coefficient of one (Agresti 2015). For zero-inflated and hurdle models, the offset is usually only included in the count part of the models, see, e.g., Lee et al. (2001), Loquiha et al. (2013), Zhen et al. (2018), and Dai et al. (2018). However, as pointed out by Feng (2022), varying exposure can also influence the probability of observing excessive zeros.

The paper's main contributions are as follows. First, we delineate several forms of hurdle-generalized Poisson (HGP) and hurdle-generalized negative binomial (HNB) regression models. It is shown that these models are useful in addressing the problem of excess zeros and over-dispersion of claim count datasets. Second, through a detailed analysis, we show that including exposure in both the zero and the count parts as a covariate is an effective approach to incorporating exposure information into zero-inflated and hurdle models. Lastly, from a practical point of view, we illustrate the effectiveness and versatility of the introduced models using real datasets and compare the results with other commonly used models.

We organize the rest of the paper as follows. Section 2 provides the mathematical background, specifically highlighting several forms of HGP and HNB regression models. Section 3 studies how to include exposure in zero-inflated and hurdle regression models. Section 4 presents real-world applications, analyzing various models using data from a Malaysian auto insurance dataset, the US National Medical Expenditure Survey, and a French auto insurance dataset. Section 5 explores the variable selection problem in the HGP and HNB models by applying the Lasso shrinkage methodology. Section 6 concludes the paper.

2. Mathematical Models

In this section, we first provide the mathematical background of generalized Poisson and generalized negative binomial models and then introduce their hurdle functional forms.

2.1. Various Forms of Generalized Poisson and Generalized Negative Binomial Random Variables

From a probability point view, the GP distribution was introduced by Consul and Jain (1973) as a limiting form of a generalized negative binomial distribution. Consul and Shoukri (1988) showed that a GP distribution can be viewed as the distribution of the number of served customers in a busy period of a queue with Poisson arrival and a constant service time. GP distribution can also be considered as the distribution of the total progeny in a Galton branching process, where both the initial number of a species and the number of offspring an individual produces follow a Poisson distribution. From a

statistical point view, the GP distribution and its related distributions are flexible and can be used to model over-dispersed or under-dispersed data.

The GP distribution has been applied in actuarial science. For instance, Gerber (1990) showed that the number of jumps it takes for a classical Poisson risk process with a constant claim size to reach a certain level follows a GP distribution. Consul (1993) compared the GP distribution with several well-known distributions and concluded that the GP distribution is a plausible model for claim frequency data. Calderín-Ojeda et al. (2019) proposed a special GP distribution, and tested the performance of their GP regression model using French Motor Personal Line datasets, which are available in the R package "CASdatasets". Scollnik (1995) presented a Bayesian analysis of GP distribution using two datasets; one was the number of injuries in automobile accidents, and the other was the ship damage incident data from Lloyd's Register of Shipping.

Different forms of GP random variables have been proposed in the literature. The classical GP-1 distribution has a probability mass function (pmf) of

$$g_1(y_i) = \mathbb{P}(Y_\mathrm{i} = y_i \mid \mu_i, a) = \frac{\mu_i(\mu_i + ay_i)^{y_i-1}}{(1+a)^{y_i} y_i!} e^{-\frac{\mu_i + ay_i}{1+a}}, \quad y_i = 0, 1, 2, \ldots,$$

where μ_i is the mean parameter and a is the dispersion parameter. The variance of GP-1 is $\mu_i(1+a)^2$. Thus, $a > 0$ implies over-dispersion, while $a < 0$ implies under-dispersion. When $a = 0$, GP-1 reduces to a Poisson distribution.

A slightly different parameterization gives the so-called GP-2 distribution with the pmf

$$g_2(y_i) = \mathbb{P}(Y_i = y_i \mid \mu_i, a) = \frac{\mu_i(\mu_i + a\mu_i y_i)^{y_i-1}}{(\mathrm{I} + a\mu_i)^{y_i} y_i!} e^{-\frac{\mu_i + a\mu_i y_i}{1+a\mu_i}}, \quad y_i = 0, 1, 2, \ldots.$$

The mean and variance of the GP-2 distribution are μ_i and $\mu_i(1+a\mu_i)^2$, respectively. While the GP-1 distribution has a linear mean–variance relationship, the GP-2 distribution has a cubic mean–variance relationship. The applications of the GP-2 distribution have been discussed in, e.g., Wang and Famoye (1997) and Ismail and Jemain (2007).

Another parameterization of the GP distribution, GP-P, which was studied in, e.g., Zamani and Ismail (2012), has the pmf

$$g_P(y_i) = \mathbb{P}(Y_i = y_i \mid \mu_i, a, P) = \frac{\mu_i\left(\mu_i + a\mu_i^{P-1} y_i\right)^{y_i-1}}{\left(1 + a\mu_i^{P-1}\right)^{y_i} y_i!} e^{-\frac{\mu_i + a\mu_i^{P-1} y_i}{1+a\mu_i^{P-1}}}, \quad y_i = 0, 1, 2, \ldots. \quad (1)$$

A GP-P random variable Y_i has mean $\mathbb{E}(Y_i) = \mu_i$ and variance $\mathrm{Var}(Y_i) = \mu_i\left(1 + a\mu_i^{P-1}\right)^2$. The additional parameter, P, provides more flexibility in modeling the variance function. It reduces to GP-1 and GP-2 regressions with $P = 1$ and $P = 2$, respectively.

The generalized negative binomial (NB-P) distribution, which was introduced in Greene (2008) and discussed in Cameron and Trivedi (2013), Hilbe (2011) and Ismail and Zamani (2013), has a parameter set (a, μ_i, P) and the pmf

$$\begin{aligned} h_P(y_i) = \mathbb{P}(Y_i = y_i \mid \mu_i, a, P) &= \frac{\Gamma\left(y_i + a^{-1}\mu_i^{2-P}\right)}{y_i!\Gamma\left(a^{-1}\mu_i^{2-P}\right)} \\ &\times \left(\frac{a^{-1}\mu_i^{2-P}}{a^{-1}\mu_i^{2-P} + \mu_i}\right)^{a^{-1}\mu_i^{2-P}} \left(\frac{\mu_i}{a^{-1}\mu_i^{2-P} + \mu_i}\right)^{y_i}, \quad y_i = 0, 1, \ldots. \end{aligned} \quad (2)$$

2.2. Hurdle Functional Form of the Generalized Poisson Regression Model

A hurdle model involves the application of two different models to analyze data that fall either above or below a specific threshold, which is typically set at zero. Therefore, it

is sometimes called a two-part model. Following Mullahy (1986), the distribution of the claim counts according to a hurdle model is given by

$$\mathbb{P}(Y_i = y_i) = \begin{cases} f_1(0), & y_i = 0, \\ \frac{1-f_1(0)}{1-f_2(0)} f_2(y_i) := \Phi f_2(y_i), & y_i = 1, 2, \ldots, \end{cases}$$

where f_1 and f_2 are two probability functions that describe the distribution of the zero and non-zero parts of Y_i. In insurance applications, the quantity Φ can be interpreted as the probability of reporting at least one claim. As argued in Boucher et al. (2007), in auto insurance, policyholders' behavior may change after a claim has been made; therefore, it is natural to apply hurdle models to describe the two parts (zero claim and non-zero claims) of the claim process. An advantage of the hurdle model is that the parameters for each part can be estimated separately.

In what follows, we assume that f_1 is Bernoulli-distributed. Then, the hurdle functional form of a generalized Poisson (HGP-P) regression model is given as

$$\mathbb{P}(Y_i = y_i) = \begin{cases} \omega_i, & y_i = 0, \\ (1-\omega_i)\frac{g_P(y_i)}{1-g_P(0)}, & y_i = 1, 2, 3, \ldots, \end{cases}$$

where $g_P(y_i)$ is defined in Equation (1). Note that the term $\frac{g_P(y_i)}{1-g_P(0)}$ is usually referred to as the zero-truncated GP distribution. In addition, we assume that μ_i is related to covariates $\mathbf{x}_i$ by a log link function

$$\log(\mu_i) = \mathbf{x}_i^T \boldsymbol{\beta}, \tag{3}$$

where $\boldsymbol{\beta}$ is the vector of regression parameters, and ω_i is related to covariates $\mathbf{z}_i$ by a logit link function

$$\log\left(\frac{\omega_i}{1-\omega_i}\right) = \mathbf{z}_i^T \boldsymbol{\gamma}. \tag{4}$$

The HGP-P model reduces to the HGP-1 and HGP-2 models when $P = 1$ and $P = 2$, respectively. Therefore, the likelihood ratio test (LRT) can be applied for testing the HGP-1 model (or HGP-2 model) against the HGP-P model.

The loglikelihood function for the HGP-P regression model is given by

$$\log L(\boldsymbol{\gamma}, \boldsymbol{\beta}, a, P) = \log L_1(\boldsymbol{\gamma}) + \log L_2(\boldsymbol{\beta}, a, P),$$

where

$$\log L_1(\boldsymbol{\gamma}) = \sum_{i=1}^{n} \left[I_{(y_i=0)} \log(\omega_i) + \left(1 - I_{(y_i=0)}\right) \log(1-\omega_i) \right],$$

and

$$\log L_2(\boldsymbol{\beta}, a, P) = \sum_{i=1}^{n} \left[1 - I_{(y_i=0)}\right] \Big\{ -\log(1-\exp(-A_i)) + (y_i - 1) \log\left(\mu_i + a\mu_i^P y_i!\right) + \log \mu_i - y_i \log\left(1 + a\mu_i^{P-1}\right) - \log(y_i!) - A_i \Big\}.$$

with $A_i = \frac{\mu_i + a\mu_i^{P-1} y_i}{1 + a\mu_i^{P-1}}$. Note that the regression parameters $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are included in the loglikelihood function through the link functions for μ_i and ω_i.

The two components of the loglikelihood function, $\log L_1(\boldsymbol{\gamma})$ and $\log L_2(\boldsymbol{\beta}, a, P)$, can be maximized separately. In particular, the parameter $\boldsymbol{\gamma}$ can be estimated using a simple logistic regression. The system of normal equations for estimating $\boldsymbol{\beta}$ is obtained by taking the partial derivative of $\log L_2(\boldsymbol{\beta}, a, P)$. Since these partial derivative equations cannot be simplified, the Newton–Raphson method is applied to solve them. The standard errors of the parameter estimates are given by the square root of the diagonal elements of the

inverse of the Hessian matrix. The estimated parameters from the truncated Poisson fit are used as starting values for faster convergence.

We note that the two-part structure of the hurdle model greatly simplifies the optimization procedure.

2.3. Hurdle Functional Form of the Generalized Negative Binomial Regression Model

The hurdle functional form of the generalized negative binomial (HNB-P) regression model is defined as

$$\mathbb{P}(Y_i = y_i) = \begin{cases} \omega_i, & y_i = 0, \\ (1-\omega_i)\frac{h_P(y_i)}{1-h_P(0)}, & y_i = 1,2,3,\ldots, \end{cases}$$

where $h_P(\cdot)$ is the NB-P pmf defined in Equation (2), ω_i is related to covariates $\mathbf{z}_i$ with a logit link function (4), and μ_i is related to covariates $\mathbf{x}_i$ via a log link function (3).

The loglikelihood function for the HNB-P regression model is given by

$$\log L(\boldsymbol{\gamma}, \boldsymbol{\beta}, a, P) = \log L_1(\boldsymbol{\gamma}) + \log L_2(\boldsymbol{\beta}, a, P),$$

where

$$\log L_1(\boldsymbol{\gamma}) = \sum_{i=1}^{n} \Big[I_{(y_i=0)} \log(\omega_i) + \Big(1 - I_{(y_i=0)}\Big) \log(1-\omega_i) \Big],$$

and

$$\begin{aligned} \log L_2(\boldsymbol{\beta}, a, P) = \sum_{i=1}^{n} \Big[1 - I_{(y_i=0)}\Big] \Big\{ & B_i \log(B_i) - y_i \log(B_i + \mu_i) - B_i \log(B_i + \mu_i) \\ & + \sum_{j=0}^{y_i - 1} \log(B_i + j) + y_i \log \mu_i - \log\big(1 - h_p(0)\big) \Big\}. \end{aligned}$$

with $B_i = a^{-1}\mu_i^{2-P}$. The estimation of the regression parameters for HNB-P is similar to that for the HGP-P model.

3. Incorporating Exposure in Zero-Inflated and Hurdle Regression Models

In many insurance loss datasets, different policyholders (observations) may have different risk exposures, yet only the total number of claims is reported. For example, a dataset could report the total number of claims made by a policyholder during the whole policy period, but different policyholders may stay in the policy for different periods of time. An offset term in the regression is a commonly used strategy for enclosing a population size at risk or the amount of exposure time. Particularly, if a log link function is used, the model can be defined as

$$\log(\mu_i) = \boldsymbol{x}_i^T \boldsymbol{\beta} + \log(E_i),$$

or equivalently $\mu_i = E_i \mathrm{e}^{x_i^T \beta}$, where E_i is the exposure for policyholder i. This approach of considering exposure makes sense because, intuitively, the mean number of events should be proportional to the size of the exposure.

For zero-inflated and hurdle models, the offset is usually only included in the count part of the models, see, e.g., Lee et al. (2001), Loquiha et al. (2013), Zhen et al. (2018), Dai et al. (2018). However, as pointed out by Feng (2022), the probability of observing excessive zeros can also be impacted by exposure in many situations. One might directly impose exposure in the zero-inflated part of the model in the same way as in the count model. For example, if the logit model is used for the zero part, we might write

$$\operatorname{logit}(\omega_i) = \log\left(\frac{\omega_i}{1-\omega_i}\right) = \boldsymbol{z}_i^T \boldsymbol{\gamma} + \log(E_i).$$

However, this may not be plausible because it indicates that the probability of zero inflation ω_i increases with the exposure size, which is counter-intuitive. Feng (2022) then proposed the model

$$\begin{aligned} \text{logit}(\omega_i) &= \boldsymbol{z}_i^T \boldsymbol{\gamma} + \xi_1 \log(E_i), \\ \log(\mu_i) &= \boldsymbol{x}_i^T \boldsymbol{\beta} + \xi_2 \log(E_i), \end{aligned} \tag{5}$$

where ξ_1 and ξ_2 are the regression coefficients for the logarithm transformed E_i. Model (5) allows risk exposures to be included in the analysis as a regular covariate in both the binary and count parts of the zero-inflated and hurdle models. We next provide a simulation study to illustrate the benefit of such a method.

A Simulation Study

In this subsection, we implement a simulation study to compare several approaches to incorporate risk exposure in zero-inflation models.

We generate 100 observations as follows. Each observation i is associated with an exposure size E_i, which is uniformly distributed among one to ten. The number of events, N_i, for the ith observation is then the summation of E_i independent and identically distributed ZIP-distributed random variables $Y^{(i)}$ with parameters (ω_i, μ_i), where ω_i is the zero-inflation probability and μ_i is the Poisson count mean. That is,

$$N_i = \sum_{j=1}^{E_i} Y_j^{(i)}.$$

Furthermore, assume that there are two covariates: $x_{1,i}$, which can take values 0 or 1, and $x_{2,i}$, which is a realization of a normal $(1, 1)$ random variable. The distribution parameters are related to the covariates by:

$$\text{logit}(\omega_i) = \gamma_0 + \gamma_1 x_{1,i} + \gamma_2 x_{2,i},$$

with $\gamma_0 = \gamma_1 = \gamma_2 = 1$, and

$$\log(\mu_i) = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i},$$

with $\beta_0 = \beta_1 = \beta_2 = 0.5$.

The mean and standard deviation of the simulated number of events are 18.4 and 32.02, respectively, where 34% of the claims are zero.

We next fit the simulated data to ZIP and ZIGP-P regression models that handle the exposures differently, as described in Equations (6)–(10).

$$\text{ZIP} \begin{cases} \text{logit}(\omega_i) = \gamma_0 + \gamma_1 x_1 + \gamma_2 x_2, \\ \log(\mu_i) = \beta_0 + \beta_1 x_1 + \beta_2 x_2, \end{cases} \tag{6}$$

$$\text{ZIP}^{ee} \begin{cases} \text{logit}(\omega_i) = \gamma_0 + \gamma_1 x_1 + \gamma_2 x_2 + \xi_1 \log(E_i), \\ \log(\mu_i) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \xi_2 \log(E_i), \end{cases} \tag{7}$$

$$\text{ZIP}^{e} \begin{cases} \text{logit}(\omega_i) = \gamma_0 + \gamma_1 x_1 + \gamma_2 x_2, \\ \log(\mu_i) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \xi_2 \log(E_i), \end{cases} \tag{8}$$

$$\text{ZIP}^{11} \begin{cases} \text{logit}(\omega_i) = \gamma_0 + \gamma_1 x_1 + \gamma_2 x_2 + \log(E_i), \\ \log(\mu_i) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \log(E_i), \end{cases} \tag{9}$$

$$\text{ZIP}^{1} \begin{cases} \text{logit}(\omega_i) = \gamma_0 + \gamma_1 x_1 + \gamma_2 x_2, \\ \log(\mu_i) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \log(E_i). \end{cases} \tag{10}$$

The parameter estimates, including absolute t-ratios, log likelihood (LL), Akaike information criterion (AIC), and Bayesian information criterion (BIC), for the ZIP model are presented in Table 1, and those for the ZIGP-P model are shown in Table 2.

Table 1. Parameter estimates, t-ratios, and model fit measures for simulated data using the ZIP model under various exposure scenarios.

	ZIP		**ZIP^{ee}**		**ZIP^{e}**		**ZIP^{11}**		**ZIP^{1}**	
Par	**Est.**	**|t.ratio|**	**Est.**	**|t.ratio|**	**Est.**	**|t.ratio|**	**Est.**	**|t.ratio|**	**Est.**	**|t.ratio|**
Logistic proportion of models										
γ_0	−1.53	34.92	0.49	7.20	−1.60	35.20	−2.92	72.02	−1.85	36.64
γ_1	0.55	12.52	0.71	14.00	0.58	12.95	0.69	17.43	0.68	14.34
γ_2	0.53	22.25	0.70	25.07	0.56	22.93	0.66	29.60	0.66	25.23
ξ_1			−1.59	39.98						
Count proportion of models										
β_0	1.86	336.0	1.16	106.9	1.16	105.6	0.04	6.75	0.04	7.51
β_1	0.80	164.4	0.79	162.5	0.79	162.5	0.79	160.9	0.79	160.8
β_2	0.83	341.6	0.82	334.4	0.82	334.4	0.81	329.1	0.81	328.9
ξ_2			0.41	78.01	0.41	78.17				
LL	−45,067		−40,652		−41,671		−50,271		−47,284	
AIC	90,147		81,319		83,356		100,553		94,579	
BIC	90,190		81,377		83,406		100,597		94,623	

Table 2. Parameter estimates, t-ratios, and model fit measures for simulated data using the ZIGP-P model under various exposure scenarios.

	ZIGP-P		**ZIGP-P^{ee}**		**ZIGP-P^{e}**		**ZIGP-P^{11}**		**ZIGP-P^{1}**	
Par	**Est.**	**|t.ratio|**	**Est.**	**|t.ratio|**	**Est.**	**|t.ratio|**	**Est.**	**|t.ratio|**	**Est.**	**|t.ratio|**
Logistic part of the models										
γ_0	−1.65	35.03	0.40	5.75	−1.74	35.16	−3.23	66.73	0.46	33.39
γ_1	0.60	13.24	0.74	14.40	0.64	13.72	0.81	18.94	−1.47	31.21
γ_2	0.58	23.24	0.74	25.52	0.62	23.94	0.77	31.09	0.61	13.49
ξ_1	–	–	−1.58	39.43						
Non-zero part of the models										
β_0	1.88	138.5	1.06	41.95	0.99	37.48	0.07	4.69	0.61	13.49
β_1	0.79	57.80	0.79	61.5	0.79	61.21	0.80	51.87	0.78	60.03
β_1	0.82	111.7	0.81	115.9	0.81	115.8	0.81	95.22	0.80	115.9
ξ_2			0.47	37.30	0.50	38.09				
a	0.39	15.31	0.17	13.41	0.19	13.28	0.45	14.50	0.39	16.60
P	1.46	75.21	1.67	77.16	1.64	75.51	1.46	72.39	1.43	82.04
LL	−30,270		−28,606		−29582		−33,088		−29,685	
AIC	60556		57,233		59,181		66,193		59,386	
BIC	60,614		57,304		59,246		66,250		59,443	

Table 1 shows that the ZIP^{ee} model has the lowest AIC and BIC values. The worst model is ZIP^{11}, which includes an offset term in the binary component and the positive count. Notice that the parameter value ξ_1 for the binary part is negative, expressing the fact that when the exposure increases, one should expect a smaller value for the zero-inflation parameter ω_i; on the other hand, the value of parameter ξ_2 for the count part is positive, expressing the fact that when the exposure increases, one should expect a greater value for the expected count μ_i. This finding highlights the importance of having exposure in both the binary and count parts of the model.

Notice that in the true model, the distribution of N_i is no longer ZIP; it is rather a summation of some random number of ZIP distributions. Therefore, there is no reason that one has to fit the data with exposure with a ZIP model.

Table 2 shows that the ZIGP-P^{ee} model fits the data better than the competing models based on AIC and BIC criteria. In addition, comparing Tables 1 and 2, we see that the ZIGP-P models perform better than the ZIP models. This is because, as discussed above,

the distribution of N_is is no longer ZIP. The ZIGP-P model, which includes two additional parameters compared to the ZIP model, presents a more flexible option for fitting the data.

Table 3 presents the average AIC and BIC values obtained from analyzing 100 simulated datasets for model comparison when the number of simulated data is 1000, 5000, and 10,000. The results demonstrate that the ZIGP-P^{ee} model performs better than the other models in all scenarios, consistently producing the smallest AIC and BIC values. These findings indicate that the ZIGP-P^{ee} model is a robust and reliable model for analyzing the simulated dataset.

Table 3. Comparing the model fitness of ZIGP-P^{ee}, ZIGP-P^{e}, ZIGP-P^{11}, and ZIGP-P^{1} based on the mean values of AIC and BIC over 100 simulated datasets created from the ZIP model.

	n	ZIGP-P^{ee}	ZIGP-P^{e}	ZIGP-P^{11}	ZIGP-P^{1}
AIC	1000	5721.15	5915.20	6633.57	6037.35
	5000	28,575.05	29,503.79	33,097.63	30,157.15
	10,000	57,194.15	59,054.38	66,010.11	60,329.38
BIC	1000	5770.23	5959.37	6672.83	6076.62
	5000	28,640.22	29,562.44	33,149.77	30,209.29
	10,000	57,266.26	59,119.27	66,067.79	60,387.06

Table 4 shows the results of fitting the HGP-P model to the simulated data in this section under different treatments of exposure. According to the AIC and BIC, the HGP-P^{ee} model outperforms other competing models, which is consistent with previous findings. This observation emphasizes the importance of including exposure (log(Exposures)) in the binary part of the HGP-P model. The estimated effect of log(Exposures) on the binary component of the HGP-P^{ee} model, $\xi_1 = -1.58$ (t-ratio = 40.38) reflects the negative association between exposure and the probability of observing an excess zero count. This finding is consistent with the ZIGP-P^{ee} model's ξ_1 estimation. In contrast, the estimated effect of log(Exposures) on the count component of the HGP-P^{ee} model is positive, with an associated effect size of $\xi_2 = 0.46$ (t-ratio = 36.68). Notably, this effect size is close to that of the ZIGP-P^{ee} model ($\xi_2 = 0.47$). Furthermore, the functional parameter for the HGP-P^{ee} model was estimated to be $P = 1.66$, which is very close to the value of $P = 1.67$ for the ZIGP-P^{ee} model.

Table 4. Parameter estimates, t-ratios, and model fit measures for simulated data using the HGP-P model under various exposure scenarios.

	HGP-P		HGP-P^{ee}		HGP-P^{e}		HGP-P^{11}		HGP-P^{1}	
Par	**Est.**	**|t.ratio|**	**Est.**	**|t.ratio|**	**Est.**	**|t.ratio|**	**Est.**	**|t.ratio|**	**Est.**	**|t.ratio|**
Logistic part of the models										
γ_0	−1.46	34.72	0.59	9.06	−1.46	34.72	−2.54	73.88	−1.46	34.72
γ_1	0.52	11.99	0.66	13.29	0.52	11.99	0.54	14.44	0.52	11.99
γ_2	0.49	21.40	0.64	24.12	0.49	21.40	0.48	24.93	0.49	21.40
ξ_1			−1.58	40.38						
Non-zero part of the models										
β_0	1.88	135.3	1.07	41.75	1.07	41.75	0.10	7.29	0.10	7.29
β_1	0.80	57.70	0.79	61.35	0.79	61.35	0.81	52.75	0.81	52.75
β_2	0.82	110.5	0.81	114.6	0.81	114.7	0.82	95.28	0.82	95.28
ξ_2			0.46	36.68	0.46	36.68				
a	0.41	14.94	0.17	13.22	0.17	13.22	0.26	15.71	0.26	15.71
P	1.45	73.14	1.66	75.85	1.66	75.85	1.62	86.18	1.62	86.18
LL	−30,287		−28,633		−29,678		−33,564		−30,472	
AIC	60,589		57,286		59,374		67,144		60,960	
BIC	60,647		57,358		59,439		67,201		61,018	

We remark that none of the models in Equations (6)–(10) "correctly" describe the underlying simulation model. Our analysis shows that models with observations with different exposures and are zero-inflated, including exposure in both the zero and the count parts as covariates (model *ee*), perform the best.

4. Model Fitting Results

In this section, we apply our proposed regression models to three datasets: the Malaysian Motor Insurance Data, the 1987/88 US National Medical Expenditure Survey data, and a French auto insurance dataset, freMTPL2freq, which is available in the R "CASdatasets" package.

4.1. Malaysian Motor Insurance Data

This dataset from Insurance Services Malaysia includes 1.01 million private car policies from ten Malaysian insurance companies in 2002. It includes information on exposures measured by the number of cars per year, claim counts for own damage and third party property damage, and four rating factors: vehicle year, vehicle make, vehicle cc, and location. The first three rating variables describe vehicle properties, whereas the last one (location) gives the location where the vehicle was operated. This dataset has been studied by Fuzi et al. (2016). As detailed therein, each of the four rating factors has five levels, amounting to $5^4 = 625$ cross-classified rating classes. Excluding 73 rating classes with zero exposure, we used 552 rating classes in this study. The response variable is the number of own damage claims in this study.

We fitted the dataset to the HP, HGP-P, and HNB-P regression models. The zero part was fit using logistic regression, and the none-zero part by maximizing the likelihood using the "nlm" function in R. This separation of the estimation of zero and non-zero parts greatly simplifies the computation.

The parameter estimates and the absolute values of the t-ratio for the models are reported in Table 5. It is seen that the over-dispersion and functional parameters (a and P) in the non-zero parts of the GP-P and NB-P models are both significant. In addition, in all models, the coefficients ξ_1 and ξ_2 for the log exposures of the zero (logistic) and the non-zero parts, respectively, are significant.

For comparison purposes, we fitted the Poisson, GP-1, GP-2, GP-P, NB-1, NB-2, NB-P, and corresponding zero-inflated and hurdle models to this dataset. The LL, AIC, and BIC for these models are provided in Table 6.

Table 5. Parameter estimates and absolute t-ratios for the Malaysian Motor Insurance Data.

	Coefficients for the Non-Zero Part of the Models						**Logistic Coef.**	
	Poisson		**GP-P**		**NB-P**			
Parameter	**Est.**	**|t.ratio|**	**Est.**	**|t.ratio|**	**Est.**	**|t.ratio|**	**Est.**	**|t.ratio|**
Intercept	−2.59	49.29	−2.75	17.14	−2.81	16.94	3.00	3.14
2–3 year	0.50	39.68	0.54	12.30	0.53	12.24	−2.14	2.92
4–5 year	0.48	36.46	0.49	10.78	0.49	10.99	−0.99	1.66
6–7 year	0.41	31.15	0.44	9.85	0.43	9.78	−1.31	2.02
above 8	0.26	20.33	0.27	6.11	0.27	6.10	0.15	0.25
1001–1300 cc	−0.10	4.40	−0.10	1.45	−0.10	1.51	−0.49	0.78
1301–1500 cc	0.10	4.26	0.07	1.12	0.08	1.07	−1.72	2.04
1501–1800 cc	0.30	12.56	0.27	3.93	0.28	3.91	−1.51	1.66
above 1800 cc	0.38	16.12	0.37	5.30	0.37	5.13	−1.47	1.49
Local type 2	−0.26	12.01	−0.33	5.31	−0.31	4.76	−0.09	0.09
Foreign type 1	−0.28	23.55	−0.25	6.14	−0.25	6.33	1.27	1.43
Foreign type 2	0.00	0.15	0.06	1.03	0.06	1.01	0.12	0.15
Foreign type 3	−0.16	7.69	−0.13	1.87	−0.13	1.90	2.03	2.03
East	0.24	13.27	0.30	5.13	0.29	4.92	−0.47	0.73
Central	0.35	30.02	0.33	8.24	0.33	8.27	−1.54	1.77
South	0.23	18.15	0.26	5.89	0.25	5.79	0.36	0.56
East Malaysia	0.08	5.48	0.07	1.42	0.08	1.53	−0.02	0.04
log(Exposure)	0.93	187.48	0.95	59.64	0.95	59.21	−1.15	6.19
a	-	-	1.51	8.01	5.34	6.64	-	-
P	-	-	1.09	42.39	1.12	34.66	-	-
LL	−3809.43		−2028.35		−2036.86		−82.35	
AIC	7654.85		4096.70		4113.73		200.71	
BIC	7730.61		4180.87		4197.90		278.35	

Based on AIC and BIC, the HGP and HNB models are obviously better than the HP model. Further, the HGP-P, HGP-1, and HNB-P models are the top three best models, followed by HNB-1. The best functional parameters in the HGP-P and HNB-P models are

$P = 1.09$ and $P = 1.12$, respectively, which are close to 1. In particular, the HNB-P model has a much lower AIC/BIC than the HNB-2 model, confirming that it is more flexible than the latter, which is accessible in the "pscl" package in R.

Table 6. The number of parameters, LL, AIC, and BIC of different models for Malaysian Motor Insurance Data.

Models	No. of Parameters	LL	AIC	BIC
Poisson	18	−3917.5	7871.1	7948.7
GP-1	19	−2166.1	4370.1	4452.1
GP-2	19	−2441.6	4921.2	5003.1
GP-P	20	−2146.4	4332.8	4419.1
NB-1	19	−2191.0	4419.9	4501.9
NB-2	19	−2324.1	4686.2	4768.2
NB-P	20	−2173.6	4387.3	4473.5
ZIP	36	−3899.9	7871.9	8027.2
ZIGP-1	37	−2281.4	4636.9	4796.5
ZIGP-2	37	−2659.2	5392.3	5551.9
ZIGP-P	38	−2167.1	4410.3	4574.2
ZINB-1	37	−2695.5	5464.9	5624.5
ZINB-2	37	−2356.8	4787.5	4947.1
ZINB-P	38	−2153.8	4383.6	4547.5
HP	36	−3891.8	7855.6	8008.9
HGP-1	37	−2116.2	4306.5	4464.1
HGP-2	37	−2420.6	4915.1	5072.7
HGP-P	38	−2110.7	4297.4	4459.2
HNB-1	37	−2125.9	4325.8	4483.4
HNB-2	37	−2321.9	4717.8	4875.4
HNB-P	38	−2119.2	4314.4	4476.2

Moreover, the coefficient for log exposure in the count part is positive, and in the logistic part it is negative. They are both significant; this verifies our simulation results in Section 3.

4.2. The US National Medical Expenditure Survey Data

We now consider the US National Medical Expenditure Survey 1987/88 data studied by Deb and Trivedi (1997). This dataset contains a subsample of 4406 observations of individuals aged 66 and over who were covered by Medicare, a public insurance program. The dataset is available from the R package accompanying Kleiber and Zeileis (2008) and is also known as "DebTrivedi". The number of physician office visits (ofp), with a mean and variance of 5.77 and 45.69, respectively, is the response variable. We fitted the data to the HP, HGP-P, and HNB-P regression models. The parameter estimates and absolute value of t-ratios are provided in Table 7. Based on the Wald test, both over-dispersion and functional parameters (a and P) are significant.

Table 7. Parameter estimates and absolute t-ratios for the US National Medical Expenditure Survey dataset.

	Coefficients for the Non-Zero Part of the Models						Logistic Coef.	
	Poisson		GP-P		NB-P			
Parameter	Est.	|t.ratio|	Est.	|t.ratio|	Est.	|t.ratio|	Est.	|t.ratio|
Intercept	1.84	20.46	1.62	7.01	1.60	6.63	−1.37	2.27
Poorhlth	0.28	15.19	0.31	6.18	0.31	6.18	0.07	0.42
Exclhlth	−0.34	10.67	−0.37	4.86	−0.39	4.81	−0.32	2.27
Numchron	0.12	24.80	0.15	12.40	0.15	11.92	0.55	12.14
Adldiff	0.12	7.19	0.10	2.26	0.12	2.71	−0.18	−1.44
Noreast	0.11	5.92	0.10	2.09	0.12	2.32	0.03	0.21
Other regions	0.02	1.24	0.01	0.33	0.02	0.46	−0.10	−0.89
Midwest	0.12	6.06	0.12	2.45	0.14	2.62	0.10	0.71
Age	−0.08	6.92	−0.07	2.47	−0.08	2.68	0.19	2.51
Black	0.00	0.03	−0.03	0.53	−0.03	0.50	−0.32	2.52
Male	−0.01	0.71	−0.02	0.65	−0.02	0.61	−0.46	4.82
Married	−0.07	4.60	−0.06	1.66	−0.07	1.80	0.25	2.41
School	0.02	9.58	0.02	3.77	0.02	3.82	0.05	4.24
Faminc	0.00	1.31	0.00	0.35	0.00	0.46	0.01	0.36
Employed	0.06	2.69	−0.01	0.10	0.03	0.56	−0.01	0.09
Private health	0.19	9.51	0.24	4.58	0.27	4.90	0.76	6.85
Medicaid	0.19	7.35	0.25	3.63	0.27	3.75	0.55	3.21
a	-	-	0.60	5.43	1.67	4.20	-	-
P	-	-	1.45	14.95	1.56	12.28	-	-

Table 8 presents the LL, AIC, and BIC for the Poisson, GP-1, GP-2, GP-P, NB-1, NB-2, NB-P, and their related zero-inflated and hurdle models. It also shows the results for some popular models used to fit the data, which include the constrained two-point finite mixture of negative binomials (CFMNB-2), the two-point finite mixture of negative binomials (FMNB-2), and the constrained three-point finite mixture of negative binomials (CFMNB-3) that were introduced by Deb and Trivedi (1997), as well as the two-point negative binomial mixture (NBM2) used by Park and Kim (2021).

Overall, the HGP-P and CFMNB-3 models, and the FMNB-2 model, which are based on NB-1 specifications, are among the preferred models according to AIC and BIC.

Table 8. Number of parameters, LL, AIC, and BIC for different models for the US National Medical Expenditure Survey dataset.

Models	No. of Parameters	LL	AIC	BIC
Poisson	17	−18,134	36,303	36,412
GP-1	18	−12,147	24,330	24,445
GP-2	18	−12,237	24,510	24,625
GP-P	19	−12,147	24,332	24,453
NB-1	18	−12,156	24,348	24,463
NB-2	18	−12,202	24,440	24,555
NB-P	19	−12,155	24,348	24,470
ZIP	34	−16,290	32,648	32,862
ZIGP-1	35	−12,096	24,261	24,485
ZIGP-2	35	−12,095	24,259	24,483
ZIGP-P	36	−12,085	24,242	24,472
ZINB-1	35	−12,133	24,336	24,560
ZINB-2	35	−12,117	24,304	24,528
ZINB-P	36	−12,114	24,301	24,531
HP	34	−16,290	32,648	32,862
HGP-1	35	−12,085	24,240	24,460
HGP-2	35	−12,096	24,262	24,482
HGP-P	36	−12,077	24,227	24,453
HNB-1	35	−12,113	24,296	24,517
HNB-2	35	−12,110	24,291	24,511
HNB-P	36	−12,104	24,280	24,507
NBM2	33	−12,139	24,343	24,554
CFMNB-2 *	21	−12,098	24,238	24,372
FMNB-2 *	37	−12,073	24,220	24,456
CFMNB-3 *	24	−12,098	24,244	24,397
CFMNB-2 **	21	−12,149	24,340	24,474
FMNB-2 **	37	−12,134	24,342	24,579
CFMNB-3 **	24	−12,149	24,346	24,499

* Based on the NB-1. ** Based on the NB-2.

4.3. The freMTPL2freq Dataset

The freMTPL2freq dataset, which is included in the "CASdatasets" package, provides information on the number of claims and risk-related features for 677,991 third party motor liability policies. Table 9 provides a summary of the covariates that were included in the analysis. The mean and variance of the number of claims are reported as 0.0532 and 0.0577, respectively. Moreover, it was observed that 94.98% of observations have zero claims.

Table 9. The description of the coviates in the French dataset.

Variable	Description
VehPower	The power of the car.
VehAge	The vehicle age in years
DriveAge	The driver age in years.
Log(density)	The log of the number of residents per square kilometer of the city where the car driver lives.
BonusMalus	Zero indicate a bonus, while one indicates a malus.
VehGas	The car's fuel equals zero for regular fuel and one for diesel.
Log(exposure)	The log of the period of exposure for a policy in years.

Table 10 compares our models with several commonly used regression models. It shows that the ZIGP-P model exhibited the lowest AIC and BIC values, indicating its

superiority in fitting the data. The ZINB-P and HGP-P models rank second and third, respectively. Furthermore, it is worth noting that the running time for the HGP-P model, thanks to its two-part model setting, is much shorter than the ZIGP-P and ZINB-P models.

Table 10. LL, AIC, BIC, and computational time (CT) for various statistical models applied to the French dataset.

Models	LL	AIC	BIC	CT (Seconds)
Poisson	−140,092	280,201	280,292	69
GP-1	−139,593	279,205	279,308	236
GP-2	−139,694	279,407	279,510	471
GP-P	−139,586	279,191	279,305	1562
NB-1	−139,602	279,222	279,325	898
NB-2	−139,700	279,419	279,521	401
NB-P	−139,596	279,212	279,327	1292
ZIP	−139,709	279,450	279,632	1850
ZIGP-1	−139,573	279,180	279,374	1711
ZIGP-2	−139,653	279,340	279,534	1331
ZIGP-P	−139,474	278,984	279,190	915
ZINB-1	−139,490	279,014	279,209	741
ZINB-2	−139,593	279,220	279,414	700
ZINB-P	−139,481	278,997	279,203	1339
HP	−139,665	279,361	279,521	125
HGP-1	−139,565	279,163	279,331	132
HGP-2	−139,572	279,177	279,345	157
HGP-P	−139,562	279,160	279,336	211
HNB-1	−139,573	279,180	279,347	139
HNB-2	−139,578	279,190	279,357	135
HNB-P	−139,571	279,178	279,354	269

Table 11 displays the estimated coefficients and absolute t-ratios for four models: ZIGP-P, ZINB-P, HGP-P, and HNB-P. In all models, both the over-dispersion parameter a and functional parameters P are statistically significant.

Table 11. Parameter estimation and absolute t-ratio for ZIGP-P, ZINB-P, HGP-P, and HNB-P models for the French dataset.

Count Model Coefficients								
	ZIGP-P		ZINB-P		HGP-P		HNB-P	
Parameter	**Est.**	**|t.ratio|**	**Est.**	**|t.ratio|**	**Est.**	**|t.ratio|**	**Est.**	**|t.ratio|**
Logistic Proportion of models								
Intercept	−0.14	0.39	−0.14	0.43	2.77	78.12	2.77	78.12
VehPower	0.10	2.51	0.10	2.67	−0.01	1.25	−0.01	1.25
VehAge	−0.06	7.61	−0.06	7.97	0.01	27.85	0.01	27.85
DrivAge	0.01	3.18	0.01	3.57	−0.01	7.83	−0.01	7.83
Log(density)	0.06	1.97	0.06	1.95	−0.03	10.56	−0.03	10.56
BonusMalus	5.93	2.54	5.93	2.38	−1.04	47.42	−1.04	47.42
VehGas	5.38	3.92	5.37	3.93	0.10	8.28	0.10	8.28
Log(Exposure)	−0.54	6.21	−0.54	6.69	−0.38	60.67	−0.38	60.67
Count Proportion of models								
Intercept	−2.45	49.70	−2.44	50.11	−5.20	8.64	−5.35	6.63
VehPower	−0.01	1.25	0.00	1.18	0.12	3.49	0.12	3.38
VehAge	−0.02	13.54	−0.02	13.79	−0.05	2.97	−0.05	2.81
DrivAge	0.00	3.96	0.00	3.70	0.01	1.00	0.01	1.04
Log(density)	0.03	7.07	0.03	7.32	0.16	3.56	0.17	2.98
BonusMalus	0.80	24.45	0.80	26.29	1.68	7.66	1.74	6.01
VehGas	−0.32	8.37	−0.32	9.74	0.31	2.05	0.31	1.96
Log(Exposure)	0.41	50.46	0.41	51.96	0.54	4.05	0.55	3.45
a	0.01	2.36	0.02	3.09	0.02	3.80	0.05	2.09
P	0.72	5.08	0.71	6.52	0.83	13.03	0.84	7.48

Further, we compared the AIC values for two situations in Table 12; the first column shows models that include exposure as an offset in the count part, and the second column shows those that include exposure as a covariate in the count and zero-inflation parts. The results indicate that the models with exposure included as a covariate in both parts have a lower AIC, suggesting that they fit the data better than those with exposure included only as an offset.

Likelihood ratio tests for various statistical models applied to the French dataset are presented in Table 13.

Table 12. Comparison of AIC values of various models with exposure included as an offset in the count part and as a covariate based on the French dataset.

Models	Exposure as an Offset in the Count Part	Exposure as a Covariate
Poisson	288,718	280,201
GP-P	287,192	279,191
NB-P	287,212	279,212
ZIP	287,774	279,450
ZIGP-P	287,102	278,984
ZINB-P	287,120	278,997
HP	284,806	279,361
HGP-P	283,571	279,160
HNB-P	283,588	279,178

Table 13. Likelihood ratio tests for various statistical models applied to the French dataset.

Models Compared	LRT Value	*p*-Value
GP-1 vs. Poisson	997.8	<0.001
GP-2 vs. Poisson	795.8	<0.001
GP-P vs. GP-1	15.6	0.0001
GP-P vs. GP-2	217.6	<0.001
NB-1 vs. Poisson	980.6	<0.001
NB-2 vs. Poisson	784.2	<0.001
NB-P vs. NB-1	11.6	0.0007
NB-P vs. NB-2	208	<0.001
ZIGP-1 vs. ZIP	272.4	<0.001
ZIGP-2 vs. ZIP	111.6	<0.001
ZIGP-P vs. ZIGP-1	196.8	<0.001
ZIGP-P vs. ZIGP-2	357.6	<0.001
ZINB-1 vs. ZIP	437.4	<0.001
ZINB-2 vs. ZIP	231.4	<0.001
ZINB-P vs. ZINB-1	19	<0.001
ZINB-P vs. ZINB-2	225	<0.001
HGP-1 vs. HP	200.3	<0.001
HGP-2 vs. HP	186.4	<0.001
HGP-P vs. HGP-1	5.1	0.024
HGP-P vs. HGP-2	19	<0.001
HNB-1 vs. HP	183.9	<0.001
HNB-2 vs. HP	173.9	<0.001
HNB-P vs. HNB-1	2.8	0.093
HNB-P vs. HNB-2	12.8	<0.001

5. The Lasso Regression

In this section, we briefly study the variable selection problem associated with the HGP-P and HNB-P regression models discussed in the paper by using the US National Medical Expenditure Survey 1987/88 data. Variable selection is important because it may simplify the regression model as well as reduce the out-of-sample prediction error.

Lasso regression, introduced in Tibshirani (1996), has been proven to be an effective method for variable selection. Park and Hastie (2007) expanded Lasso regression to a generalized linear model to handle count data. Related to this paper's context, Tang et al. (2014) proposed an EM adaptive Lasso method to select risk factors (covariates) for an auto insurance claim dataset. Wang et al. (2015) employed it to address the issue of variable selection for a model with zero inflation and over-dispersion.

In this study, we apply a simplified version of the Lasso shrinkage method, which aims to maximize the penalized log likelihood function

$$\log L - \lambda \sum_{i \geq 1} |\beta_i|,$$

where $\lambda \geq 0$ is the tuning parameter and β_i are the parameters of interest. The intercept β_0 and the model parameters a and P are excluded from the penalty.

When λ increases, the estimates of the coefficient values deviate from maximum likelihood estimates, resulting in lower in-sample goodness-of-fit. However, the model is simplified, potentially improving the out-of-sample performance.

Since the LL of the logistic and truncated parts of hurdle models can be separated, we may perform Lasso regression separately for the two parts. Lasso regression for the logistic part can be executed in R utilizing the "glmnet" package. Lasso regression for the truncated

functional form of generalized Poisson (TGP-P) regression and the truncated functional form of generalized negative binomial (TNB-P) regression has not been implemented in the literature. Therefore, it is implemented based on our own R codes.

To obtain the optimal value of λ that leads to the most accurate out-of-sample prediction, we applied five-fold cross-validation.

As shown in Table 14, for the logistics parts, we find that the tuning parameter is 10; four variables, "regionnortheast", "age", "faminc" and "employedye", are removed from the models. This results in a decrease in the out-of-sample deviance from 746.0 to 740.4.

Table 14. Modeling results for the original full logistic regression model and shrunken model applied to the US National Medical Expenditure Survey dataset.

	Full Model		**Lasso Regression**	
Variables	**Est.**	***p*-Value**	**Est.**	***p*-Value**
Intercept	−1.04	0.00	−1.24	0.00
healthpoor	−0.53	0.00	−0.28	0.09
healthexcellent	0.62	0.00	0.36	0.03
numchron	0.09	0.10	0.02	0.76
adldiffyes	−0.20	0.17	−0.08	0.56
regionnoreast	−0.04	0.80	0.00	1.00
regionother	0.12	0.35	0.07	0.53
regionwest	−0.30	0.06	−0.13	0.33
age	0.01	0.78	0.00	1.00
blackyes	0.57	0.00	0.40	0.00
gendermale	0.52	0.00	0.39	0.00
marriedyes	−0.24	0.03	−0.08	0.46
school	0.14	0.01	0.10	0.05
faminc	0.02	0.71	0.00	1.00
employedyes	0.04	0.80	0.00	1.00
privinsyes	−1.02	0.00	−0.82	0.00
medicaidyes	−0.57	0.00	−0.22	0.23
In-sample LL	−1436.9		−1446.7	
Out-of-sample LL	−373.0		−370.2	
In-sample deviance	2873.8		2893.4	
Out-of-sample deviance	746.0		740.4	

The results of the Lasso regression with TNB-P and TGP-P models are shown in Table 15. At the optimal value of the tuning parameter λ (18.95 and 10.77 for TNB-P and TGP-P, respectively), shrunken models lead to lower out-of-sample deviances and thus perform better than the full models. Furthermore, the out-of-sample prediction accuracy of the TNB-P model is lower than that of the TGP-P model.

Table 15. Modeling results for the original full TGP-P and TNB-P regression and shrunken models applied to the US National Medical Expenditure Survey dataset.

	TGP-P				**TNB-P**			
	Full Model		**Lasso Reg.**		**Full Model**		**Lasso Reg.**	
Variables	**Est.**	***p*-Val**	**Est.**	***p*-Val**	**Est.**	***p*-Val**	**Est.**	***p*-Val**
Intercept	1.27	0.00	1.39	0.00	1.19	0.00	1.30	0.00
healthpoor	0.31	0.00	0.26	0.00	0.32	0.00	0.30	0.00
healthexcellent	−0.42	0.00	−0.28	0.00	−0.40	0.00	−0.31	0.00
numchron	0.15	0.00	0.15	0.00	0.14	0.00	0.14	0.00
adldiffyes	0.10	0.04	0.07	0.14	0.12	0.02	0.11	0.04
regionnoreast [a]	0.13	0.02	0.06	0.23	0.11	0.06	0.04	0.46
regionother [b,c]	0.04	0.45	0.00	1.00	0.06	0.23	0.00	1.00
regionwest	0.12	0.03	0.05	0.32	0.14	0.02	0.07	0.20
age [a]	−0.04	0.04	−0.03	0.14	−0.04	0.05	−0.04	0.09
blackyes [b,c]	−0.03	0.50	0.00	1.00	0.01	0.92	0.00	1.00
gendermale	−0.03	0.53	−0.01	0.72	−0.05	0.27	−0.04	0.34
marriedyes	−0.08	0.08	−0.05	0.27	−0.05	0.26	−0.04	0.41
school	0.07	0.00	0.06	0.00	0.09	0.00	0.08	0.00
faminc [a,b]	−0.01	0.76	0.00	1.00	−0.02	0.34	−0.02	0.43
employedyes [a,b,c]	0.10	0.12	0.00	1.00	0.07	0.32	0.00	1.00
privinsyes	0.24	0.00	0.15	0.01	0.29	0.00	0.22	0.00

Table 15. *Cont.*

	TGP-P				TNB-P			
	Full Model		Lasso Reg.		Full Model		Lasso Reg.	
Variables	**Est.**	***p*-Val**	**Est.**	***p*-Val**	**Est.**	***p*-Val**	**Est.**	***p*-Val**
medicaidyes	0.27	0.00	0.16	0.04	0.30	0.00	0.24	0.00
a	0.57	0.00	0.67	0.00	1.85	0.00	2.00	0.00
P	1.48	0.00	1.40	0.00	1.49	0.00	1.45	0.00
In-sample LL	−8290.6		−8297.7		−8311.3		−8314.3	
Out-of-sample LL	−2108.3		−2105.1		−2106.7		−2088.2	
In-sample deviance	2825.5		2852.8		2734.2		2748.4	
Out-of-sample deviance	762.3		759.0		755.9		720.8	

[a] removed variable based on logistic Lasso. [b] removed variable based on TGP-P Lasso. [c] removed variable based on TNB-P Lasso.

Considering both Tables 14 and 15, we can see that the "employedyes" should be removed from the zero and non-zero parts of the model. However, other variables that were candidates for removal in zero and non-zero parts are different.

6. Discussion and Conclusions

In this paper, we explored the zero-inflated and hurdle-generalized Poisson/negative binomial models for analyzing count data. It was shown that such models can effectively tackle the common challenges of excessive zero and over-dispersion in analyzing insurance claim data. The nested structure of the models mentioned allows for the use of a likelihood ratio test to select the most appropriate model.

Further, we provided a detailed study of how to include exposure information in zero-inflated and hurdle models. We find that including exposure as a covariate in both the zero and non-zero parts can provide superior results than just including it in the non-zero part as an offset.

Finally, we showed that Lasso regression can be applied to HGP-P and HNB-P regression models for variable selection.

There are several directions to be explored for future research. One is to apply Bayesian methods to GP regression models, focusing on modeling over-dispersed count data. An earlier study in this direction was presented by Scollnik (1995). Another direction is to investigate the variable selection of zero-inflated or hurdle models. Techniques such as linear shrinkage, pretest, shrinkage pretest, Stein-type, and positive Stein-type Liu estimators, see, e.g., Stein (1981), Ledoit and Wolf (2003), and Månsson et al. (2012), could be considered in the context of the ZIGP-P or HGP-P models.

Author Contributions: Conceptualization, P.F., S.L. and J.R.; methodology, P.F., S.L. and J.R.; software, P.F.; validation, P.F., S.L. and J.R.; formal analysis, P.F., S.L. and J.R.; investigation, P.F., S.L. and J.R.; resources, P.F.; data curation, P.F.; writing—original draft preparation, P.F.; writing—review and editing, P.F., S.L. and J.R.; All authors have read and agreed to the published version of the manuscript.

Funding: Shu Li was funded by the Natural Sciences and Engineering Research Council of Canada (NSERC), grant number RGPIN-2019-06219. Jiandong Ren was funded by NSERC grant number RGPIN-2019-06561.

Data Availability Statement: In support of the results reported in this paper, the French Motor Third Party Liability Claims dataset, freMTPL2freq, was utilized and can be accessed through the 'CASdatasets' package in R. This dataset can be loaded directly using the library (CASdatasets) and data (freMTPL2freq) commands in R. Additionally, the US National Medical Expenditure Survey dataset was also employed, and is available via the 'MixAll' library in R, accessible with the commands library (MixAll) and data (DebTrivedi). It is important to highlight that, due to privacy and ethical constraints, the Malaysian dataset referenced in this study is not available for public sharing. We adhere strictly to MDPI's data-sharing policies, ensuring that all data supporting our findings, except those restricted, are readily accessible. Detailed guidelines and policies regarding data availability are available on the MDPI Ethics website.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

A full list of the abbreviations used in this manuscript (ordered alphabetically):

AIC	Akaike information criterion
BIC	Bayesian information criterion
CFMNB-2	Constrained two-point finite mixture of negative binomials
CFMNB-3	Constrained three-point finite mixture of negative binomials
FMNB-2	Two-point finite mixture of negative binomials
LL	Log likelihood
LRT	Likelihood ratio test
GP	Generalized Poisson
GP-P	Functional form of generalized Poisson
HGP	Hurdle-generalized Poisson
HGP-P	Hurdle functional form of generalized Poisson
HNB	Hurdle negative binomial
HNB-P	Hurdle functional form of negative binomial
HP	Hurdle Poisson
NB-P	Functional form of negative binomial
NBM2	Two-point negative binomial mixture
TGP-P	Truncated functional form of generalized Poisson
TNB-P	Truncated functional form of negative binomial
ZIGP-P	Zero-inflated functional form of generalized Poisson
ZINB-P	Zero-inflated functional form of negative binomial
ZIP	Zero-inflated Poisson

References

Agresti, Alan. 2015. *Foundations of Linear and Generalized Linear Models*. Hoboken: John Wiley & Sons.

Bhaktha, Nivedita. 2018. Properties of Hurdle Negative Binomial Models for Zero-Inflated and Overdispersed Count Data. Ph.D. Thesis, The Ohio State University, Columbus, OH, USA.

Boucher, Jean-Philippe, Michel Denuit, and Montserrat Guillén. 2007. Risk classification for claim counts: A comparative analysis of various zero inflated mixed Poisson and hurdle models. *North American Actuarial Journal* 11: 110–31. [CrossRef]

Calderín-Ojeda, Enrique, Emilio GóMez-Déniz, and Inmaculada Barranco-Chamorro. 2019. Modelling zero-inflated count data with a special case of the generalised Poisson distribution. *ASTIN Bulletin: The Journal of the IAA* 49: 689–707. [CrossRef]

Cameron, A. Colin, and Pravin K. Trivedi. 2013. *Regression Analysis of Count Data*. Cambridge: Cambridge University Press, vol. 53.

Consul, Prem. C. 1993. A model for distributions of injuries in auto-accidents. *Insurance: Mathematics and Economics* 13: 147. [CrossRef]

Consul, Prem. C., and Mohamed M. Shoukri. 1988. Some chance mechanisms related to a generalized poisson probability model. *American Journal of Mathematical and Management Sciences* 8: 181–202. [CrossRef]

Consul, Prem C., and Gaurav C. Jain. 1973. A generalization of the Poisson distribution. *Technometrics* 15: 791–9. [CrossRef]

Cragg, John G. 1971. Some statistical models for limited dependent variables with application to the demand for durable goods. *Econometrica: Journal of the Econometric Society* 39: 829–44. [CrossRef]

Czado, Claudia, Vinzenz Erhardt, Aleksey Min, and Stefan Wagner. 2007. Zero-inflated generalized Poisson models with regression effects on the mean, dispersion and zero-inflation level applied to patent outsourcing rates. *Statistical Modelling* 7: 125–53. [CrossRef]

Dai, Lin, Michael D. Sweat, and Mulugeta Gebregziabher. 2018. Modeling excess zeros and heterogeneity in count data from a complex survey design with application to the demographic health survey in sub-saharan africa. *Statistical Methods in Medical Research* 27: 208–20. [CrossRef]

Dean, Charmaine, Jerry. F. Lawless, and Gord. E. Willmot. 1989. A mixed Poisson–inverse-gaussian regression model. *Canadian Journal of Statistics* 17: 171–81. [CrossRef]

Deb, Partha, and Pravin K. Trivedi. 1997. Demand for medical care by the elderly: A finite mixture approach. *Journal of Applied Econometrics* 12: 313–336. [CrossRef]

Denuit, Michel, Xavier Maréchal, Sandra Pitrebois, and Jean-François Walhin. 2007. *Actuarial Modelling of Claim Counts: Risk Classification, Credibility and Bonus-Malus Systems*. Hoboken: John Wiley & Sons.

Dionne, Georges, and Charles Vanasse. 1989. A generalization of automobile insurance rating models: The negative binomial distribution with a regression component. *ASTIN Bulletin: The Journal of the IAA* 19: 199–212. [CrossRef]

Famoye, Felix, and Karan P. Singh. 2006. Zero-inflated generalized poisson regression model with an application to domestic violence data. *Journal of Data Science* 4: 117–30. [CrossRef]

Feng, Cindy. 2022. Zero-inflated models for adjusting varying exposures: A cautionary note on the pitfalls of using offset. *Journal of Applied Statistics* 49: 1–23. [CrossRef]

Frees, Edward W. 2009. *Regression Modeling with Actuarial and Financial Applications*. Cambridge: Cambridge University Press.

Frees, Edward W., Richard A. Derrig, and Glenn Meyers. 2014. *Predictive Modeling Applications in Actuarial Science*. Cambridge: Cambridge University Press, vol. 1.

Frees, Edward W., and Emiliano A. Valdez. 2008. Hierarchical insurance claims modeling. *Journal of the American Statistical Association* 103: 1457–69. [CrossRef]

Fuzi, Mohd Fadzli Mohd, Abdul Aziz Jemain, and Noriszura Ismail. 2016. Bayesian quantile regression model for claim count data. *Insurance: Mathematics and Economics* 66: 124–37. [CrossRef]

Gerber, Hans U. 1990. When does the surplus reach a given target? *Insurance: Mathematics and Economics* 9: 115–9. [CrossRef]
Greene, William. 2008. Functional forms for the negative binomial model for count data. *Economics Letters* 99: 585–90. [CrossRef]
Hilbe, Joseph M. 2011. *Negative Binomial Regression*. Cambridge: Cambridge University Press.
Ismail, Noriszura, and Abdul Aziz Jemain. 2007. Handling overdispersion with negative binomial and generalized poisson regression models. In *Casualty Actuarial Society Forum*. Arlington County: Casualty Actuarial Society, vol. 2007, pp. 103–58.
Ismail, Noriszura, and Hossein Zamani. 2013. Estimation of claim count data using negative binomial, generalized Poisson, zero-inflated negative binomial and zero-inflated generalized Poisson regression models. In *Casualty Actuarial Society E-Forum*. Arlington County: Casualty Actuarial Society, vol. 41, pp. 1–28.
Kleiber, Christian, and Achim Zeileis. 2008. *Applied Econometrics with R*. Berlin: Springer Science & Business Media.
Lambert, Diane. 1992. Zero-inflated poisson regression, with an application to defects in manufacturing. *Technometrics* 34: 1–14. [CrossRef]
Ledoit, Olivier, and Michael Wolf. 2003. Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *Journal of Empirical Finance* 10: 603–21. [CrossRef]
Lee, Andy H, Kui Wang, and Kelvin K. W. Yau. 2001. Analysis of zero-inflated Poisson data incorporating extent of exposure. *Biometrical Journal* 43: 963–75. [CrossRef]
Loquiha, Osvaldo, Niel Hens, Leonardo Chavane, Marleen Temmerman, and Marc Aerts. 2013. Modeling heterogeneity for count data: A study of maternal mortality in health facilities in mozambique. *Biometrical Journal* 55: 647–60. [CrossRef] [PubMed]
Månsson, Kristofer, B. M. Golam Kibria, and Ghazi Shukur. 2012. On liu estimators for the logit regression model. *Economic Modelling* 29: 1483–88. [CrossRef]
Mullahy, John. 1986. Specification and testing of some modified count data models. *Journal of Econometrics* 33: 341–65. [CrossRef]
Park, Myung Hyun, and Joseph H. T. Kim. 2021. Modelling healthcare demand count data with excessive zeros and overdispersion. *Global Economic Review* 50: 358–81. [CrossRef]
Park, Mee Young, and Trevor Hastie. 2007. L1-regularization path algorithm for generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 69: 659–77. [CrossRef]
Renshaw, Arthur E. 1994. Modelling the claims process in the presence of covariates. *ASTIN Bulletin: The Journal of the IAA* 24: 265–85. [CrossRef]
Saffari, Seyed Ehsan, Robiah Adnan, and William Greene. 2013. Investigating the impact of excess zeros on hurdle-generalized Poisson regression model with right censored count data. *Statistica Neerlandica* 67: 67–80. [CrossRef]
Scollnik, David P. M. 1995. Bayesian analysis of two overdispersed Poisson models. *Biometrics* 51: 1117–26. [CrossRef]
Stein, Charles M. 1981. Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics* 9: 1135–51. [CrossRef]
Tang, Yanlin, Liya Xiang, and Zhongyi Zhu. 2014. Risk factor selection in rate making: EM adaptive LASSO for zero-inflated poisson regression models. *Risk Analysis* 34: 1112–27. [CrossRef]
Tibshirani, Robert. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)* 58: 267–88. [CrossRef]
Wang, Shuo, Wangxue Chen, Meng Chen, and Yawen Zhou. 2023. Maximum likelihood estimation of the parameters of the inverse gaussian distribution using maximum rank set sampling with unequal samples. *Mathematical Population Studies* 30: 1–21. [CrossRef]
Wang, Weiren, and Felix Famoye. 1997. Modeling household fertility decisions with generalized Poisson regression. *Journal of Population Economics* 10: 273–83. [CrossRef] [PubMed]
Wang, Zhu, Shuangge Ma, and Ching-Yun Wang. 2015. Variable selection for zero-inflated and overdispersed data with application to health care demand in germany. *Biometrical Journal* 57: 867–84. [CrossRef]
Wüthrich, Mario V., and Michael Merz. 2008. *Stochastic Claims Reserving Methods in Insurance*. Hoboken: John Wiley & Sons.
Yip, Karen C. H., and Kelvin K. W. Yau. 2005. On modeling claim frequency data in general insurance with extra zeros. *Insurance: Mathematics and Economics* 36: 153–63. [CrossRef]
Zamani, Hossein, and Noriszura Ismail. 2012. Functional form for the generalized poisson regression model. *Communications in Statistics-Theory and Methods* 41: 3666–75. [CrossRef]
Zhen, Zhen, Liyang Shao, and Lianjun Zhang. 2018. Spatial hurdle models for predicting the number of children with lead poisoning. *International Journal of Environmental Research and Public Health* 15: 1792. [CrossRef]
Zuo, Guoxin, Kang Fu, Xianhua Dai, and Liwei Zhang. 2021. Generalized Poisson hurdle model for count data and its application in ear disease. *Entropy* 23: 1206. [CrossRef]