http://communications.science.ankara.edu.tr

Commun.Fac.Sci.Univ.Ank.Ser. A1 Math. Stat.
Volume 71, Number 2, Pages 601–615 (2022)
DOI:10.31801/cfsuasmas.988880
ISSN 1303-5991 E-ISSN 2618-6470

Research Article; Received: August 31, 2021; Accepted: December 14, 2021

POISSON AND NEGATIVE BINOMIAL REGRESSION MODELS FOR ZERO-INFLATED DATA: AN EXPERIMENTAL STUDY

Gizem YILDIRIM,[1] Selahattin KACIRANLAR,[2] and Hasan YILDIRIM[3]

[1]Social Security Institution, 70200 Karaman, TURKEY
[2]Department of Statistics, Çukurova University, 01330 Adana, TURKEY
[3]Department of Mathematics, Karamanoğlu Mehmetbey University,
70100 Karaman, TURKEY

Abstract. Count data regression has been widely used in various disciplines, particularly health area. Classical models like Poisson and negative binomial regression may not provide reasonable performance in the presence of excessive zeros and overdispersion problems. Zero-inflated and Hurdle variants of these models can be a remedy for dealing with these problems. As well as zero-inflated and Hurdle models, alternatives based on some biased estimators like ridge and Liu may improve the performance against to multicollinearity problem except excessive zeros and overdispersion. In this study, ten different regression models including classical Poisson and negative binomial regression with their variants based on zero-inflated, Hurdle, ridge and Liu approaches have been compared by using a health data. Some criteria including Akaike information criterion, log-likelihood value, mean squared error and mean absolute error have been used to investigate the performance of models. The results show that the zero-inflated negative binomial regression model provides the best fit for the data. The final model estimations have been obtained via this model and interpreted in detail. Finally, the experimental results suggested that models except the classical models should be considered as powerful alternatives for modelling count and give better insights to the researchers in applying statistics on working similar data structures.

2020 *Mathematics Subject Classification.* 62J05, 62J07, 62J12.
Keywords. Poisson regression, negative binomial regression, zero inflated regression, Hurdle regression, ridge regression, Liu regression.

girayg@outlook.com-Corresponding author; 0000-0002-1333-788X
skacir@cu.edu.tr; 0000-0003-0678-7935
hasanyildirim@kmu.edu.tr; 0000-0003-4582-9018

.

1. Introduction

From past to present, due to it's interpretability and simplicity abilities, regression analysis have been widely used in various fields. Particularly, it has been more attractive in machine learning field as a result of processing increasing data in this information era. The aim of a classical regression is to establish a mathematical model between a response variable and a group of explanatory variables to get some insights and some inferences. The structure of response variable is critical to determine an appropriate regression model. Although a continuous response variables is the most common one in regression models, there has been great interest to develop models on discrete variables. Count data regression is an effective way for this purpose. When the response variable is the count of some occurrences of an event and non-negative valued, count data regression models can produce convincing results. Classical regression models are based on some assumptions like normality, linearity and homoskedasticity. These assumptions are adversely affected in the existence of count response variable. In a such model's results may be to biased, unstable and poor on generalization.

In order to more suitable models for count data, some alternative models based on poisson and negative binomial distributions have been proposed. Poisson regression (PR) and negative binomial regression (NBR) models have been used with great interest in many areas like medicine, biostatistics, biology, finance, demography, astronomy, business and management, earth sciences, communication and insurance [1,2]. The underlying remarks of attention to count data regression can be given as follows: (1) interpretability, (2) easy-implementation, (3)good performance and (4) wide application area. However, it is not rare to face excessive zero values meaning no-occurrence in the variable of interest. In such situation, classical PRR and NBR models may not be sufficiently enough for count data modelling. For example, the number of cigarettes smoked daily or the number of weekly deaths due to the cancer in a hospital may be zero. The excess zeros cause a problem calles overdispersion which has the effect on increasing the sample variance [3]. Due to the overdispersion,the conditional variance could be bigger than conditional mean unlike the assumption based on the equality of them. That's why, the effect of overdispersion over classical poisson regression is more severe than negative binomial regression. Even NBR can be used to a certain extent in the existence of excessive zeros, zero inflated poisson regression (ZPR), zero inflated negative binomial regression (ZNBR), poisson Hurdle regression (PHR) and negative binomial Hurdle regression (NBHR) models have been proposed to deal with overdispersion problem. On the other side, a phenomenon call multicollinearity corresponding high correlations between two or more explanatory variables is another common problem in real word applications. In count data regression models, multicollinearity affected the significance of each variable and the stability performances. As a solution to multicollinearity in classical linear regression models, ridge regression estimator and Liu estimator were proposed by Hoerl and Kennard [4] and Liu [5], respectively. The superiorities of ridge estimator have been extended to count data regression via poisson ridge regression (PRR) and negative binomial ridge regression (NBRR) models which were proposed by Månsson and Shukur [6] and Månsson [7], respectively. Similarly, poisson Liu regression (PLR) and negative binomial Liu regression (NBLR) models based on Liu estimator were developed by Månsson and Kibria [8] and Månsson [9], respectively.

Both classical poisson and negative binomial regression models and the corresponding zero-inflated versions have been extensively used in real word applications. Wang and Famoye [10] were used the generalized poisson regression (GPR) and classical poisson and negative binomial regression models to investigate household fertility decisions. Famoye and Singh [11] proposed zero-inflated generalized poisson regression (ZGPR) model to make predictions on domestic violence data set and found this model adequate over its competitors like PR, GPR, ZPR and ZNBR. Bandyopadhyay et al. [12] used Hurdle and zero-inflated regression models on drug addiction data set. Similarly, Buu et al. [13] developed a new variable selection method for ZPR on a data set related with alcohol addiction. Mouatassim and Ezzahid [14] applied PR and ZPR models to make estimation on private health insurance data set. Xie et al. [15] made a comparison between zero-inflated models and NBR model by using a smoking data set and found that NBR model could provide convincing estimations for predicting zero-excessive data. Liyanage et al. [16] used PR model to estimate the prevalence of end-stage kidney disease and worldwide use of renal replacement therapy (RRT) and made some projections about the needs to 2030. Martinez et al. [17] compared PR and NBR models performance on a data set related with severe chronic obstructive pulmonary disease. Oliveira et al. [18] presented a comparative study on the usage of ZPR and ZNBR models for radiation-induced chromosome aberration data and developed a score test for ZPR model. Tang et al. [19] studied on zero-and-one-inflated poisson models, proposed a sampling approach with a simulation study and evaluated the performance via two real data sets. Chai et al. [20] used ZNBR model to estimate ship sinking accident mortalities. This study investigated the effective factors on medical visit and the performance of regression models on it's estimation. Ten different regression models including PR, NBR, ZPR, ZNBR, PHR, NBHR, PRR, PLR, NBRR and NBLR have been considered in evaluation. The data set is obtained from Deb and Trivedi's study [21].

The rest of this study is organized as follows. The background methodology is reviewed in Section 2. The appropriate selection methods of ridge and Liu biasing parameters are presented in Section 3. In Section 4, some information about data set are described. Performance evaluation is given in Section 5. Some conclusion and discussions are summarized in Section 6.

2. Backround Methodology

In this section, we review the aforementioned models briefly. Firstly, we give models related with conventional poisson regression including PR, PRR, PLR, ZPR and PHR. Secondly, models based on negative binomial regression including NBR, NBRR, NBLR, ZNBR and NBHR are presented.

2.1. **Models based on Poisson Regression.** Poisson regression models have been widely used and popular in various fields due to its advantages like the interpretability and suitability on inference count data. The conventional PR models are based on poisson distribution which is given as follows:

$$f\left(y_i|x_i\right)=\frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!},\ y=0,1,2,...$$

where μ is the mean parameter defined by $E\left[y_i|x_i\right]=\mu_i=\exp\left(x_i^{'}\beta\right)$, x_i corresponds the ith row of $X_{n\times(p+1)}$ data matrix, $\beta_{(p+1)\times 1}$ is the coefficients vector and y is the response variable. Poisson log likelihood function is defined as follows:

$$L\left(\mu,y\right)=\sum_{i=1}^{n}\left\{y_i\ln\left(\mu_i\right)-\mu_i-\ln\left(y_i!\right)\right\}$$

Then, poisson log likelihood function can be expressed by considering $\mu_i=\exp\left(x_i^{'}\beta\right)$ equation as follows:

$$L\left(\beta,y\right)=\sum_{i=1}^{n}\left\{y_i\left(x_i^{'}\beta\right)-\exp\left(x_i^{'}\beta\right)-\ln\left(y_i!\right)\right\}$$

The first derivative or gradient vector of $L\left(.\right)$ likelihood function can be obtained as

$$\frac{\partial\left(L\left(\beta,y\right)\right)}{\partial\beta}=\sum_{i=1}^{n}\left(y_i-\exp\left(x_i^{'}\beta\right)\right)x_i^{'}$$

An iterative algorithm like iteratively reweighted least squares [22] can be used to get the solutions of above equation. Based on IRLS algorithm the maximum likelihood estimations of β can be given as follows:

$$\hat{\beta}_{PR}=\left(X^{'}\hat{W}_{IRLS-PR}X\right)^{-1}X^{'}\hat{W}_{IRLS-PR}\hat{z}$$

where
$\hat{W}_{IRLS-PR}=\text{diag}\left[\hat{\mu}_i\left(\hat{\beta}_{PR}\right)\right]$ and $z_i=\log\left(\hat{\mu}_i\left(\hat{\beta}_{PR}\right)\right)+\left(y_i-\hat{\mu}_i\left(\hat{\beta}_{PR}\right)\right)/\hat{\mu}_i\left(\hat{\beta}_{PR}\right)$.
The steps of IRLS are repeated until a converge criterion is met. Each of $\hat{W}$ and z parameters is updated to maximize likelihood function. The threshold for converge is usually determined as 10^{-6} [23]. The matrix $\left(X^{'}\hat{W}_{IRLS-PR}X\right)$ is adversely affected due to multicollinearity. In other words, when the explanatory variables

are highly correlated, the deviations of coefficients will be unstable and the estimation variance increases. Similar to classical linear regression, ridge regression estimator can be integrated to poisson regression to deal with this issue. Månsson and Shukur [6] proposed PRR method which considers ridge regression estimator in poisson regression estimation process to deal with multicollinearity and defined as follows:

$$\hat{\beta}_{PRR} = \left(X^{'}\hat{W}_{IRLS-PR}X + kI\right)^{-1} X^{'}\hat{W}_{IRLS-PR}X\hat{\beta}_{PR},\ k \geq 0$$

As an alternative to ridge regression, Liu estimator which was proposed by Liu [5] can compete with ridge estimator for dealing with multicollinearity. One superiority of Liu estimator over ridge is to have a linear form of biasing parameters unlike non-linear form in ridge estimator. That's why, PLR was proposed by Månsson et al. [8] as follows:

$$\hat{\beta}_{PLR} = \left(X^{'}\hat{W}_{IRLS-PR}X + I\right)^{-1} \left(X^{'}\hat{W}_{IRLS-PR}X + dI\right)\hat{\beta}_{PR},\ 0 < d < 1$$

As well as multicollinearity, overdispersion is a critical problem on the performance of models based on poisson regression. The main reason of this problem is to observe excessive zeros in data set and this situation is not rare in practical real applications. The positive count values can be estimated more accurate via poisson distribution but zero counts cause zero-inflation on the model. To deal with overdispersion, zero-inflated and Hurdle regression models have been proposed. These models account excessive zeros separately in estimation process. The modelling stage consist two parts, one for estimation of positive counts and the other one for estimation of zero count values. Generally, while logistic or probit model is used to estimate zero values in a binary logic, classical poisson distribution is used for positive count values. Although zero inflated and Hurdle models have similar estimation process, there are some differences between them. Firstly, Hurdle models consider zero values independently from count values and creates a different process (i.e. different distribution) generating zeros. So, only one process can produce zero count values. At the end of estimation process, the likelihood function is calculated by using the mixture of different distributions. In zero-inflated models, the observation is possible in each of two processes. Unlike Hurdle models, the process used in estimation count values can produce zero values. Secondly, the distribution used in second part of estimation is the truncated version of conventional distribution. For example, the truncated poisson distribution is considered in Hurdle regression and non-zero count estimations are guaranteed in this way. ZPR model was proposed by Lambert [24]. The assumption underlying ZPR model is expressed as follows:

$$y_i \sim \left\{\begin{matrix} 0, with\ probability\ 1-\alpha_i \\ Poisson\left(\mu_i\right),\ with\ probability\ \alpha_i \end{matrix}\right\}$$

The unconditional probability distribution of ZPR model can be given as

$$P_{ZPR}(y_i) = \left\{ \begin{matrix} \alpha_i + (1-\alpha_i)e^{-\mu_i},\ y_i = 0 \\ (1-\alpha_i)\frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!},\ y_i \geq 1;\ 0 \leq \alpha_i \leq 1 \end{matrix} \right\}$$

Similar to ZPR models, the distribution of ZHR model is defined as follows:

$$P_{ZHR}(y_i) = \left\{ \begin{matrix} (1-\alpha_i),\ y_i = 0 \\ \alpha_i \frac{f(j;\mu_i)}{1-f(0;\mu_i)},\ y_i = j;\ j = 1, 2, ... \end{matrix} \right\}$$

where $f(j; \mu_i)$ is a truncated poisson distribution.

2.2. **Models based on Negative Binomial Regression.** Negative binomial regression model is a powerful tool which is highly effective on a wide range of count data applications. Various types of NBR model have been described. The most well-known and used type is termed as NB2 model referring quadratic form of variance function. Unlike to conventional PR model, the conventional NBR model can deal with over-dispersion problem. NBR model accounts over-dispersion by adding an additional term into probability function and this function is defined as follows:

$$P_{NBR}(y_i) = \left\{ \frac{\Gamma\left(y_i + 1/\theta\right)}{\Gamma\left(y_i+1\right)\Gamma\left(1/\theta\right)} \left(\frac{1}{1+\theta\mu_i}\right)^{1/\theta} \left(1 - \frac{1}{1+\theta\mu_i}\right)^{y_i} \right\}$$

where θ is the overdispersion parameter. It is clear that θ is becomes zero, NBR model will be equivalent to PR model. The log likelihood function of NBR model can be expressed as

$$L\left(\mu, y, \theta\right) = \sum_{i=1}^{n} y_i \ln\left(\frac{\theta\mu_i}{1+\theta\mu_i}\right) - \frac{1}{\theta}\ln\left(1+\theta\mu_i\right) + \ln\Gamma\left(y_i + \frac{1}{\theta}\right) - \ln\Gamma\left(y_i+1\right) - \ln\Gamma\left(\frac{1}{\theta}\right)$$

This likelihood function can be re-expressed in terms of coefficients vector as follows:

$$L\left(\beta_j, y, \theta\right) = \sum_{i=1}^{n} y_i \ln\left(\frac{\theta\exp\left(x_i^{'}\beta\right)}{1+\theta\exp\left(x_i^{'}\beta\right)}\right) - \frac{1}{\theta}\ln\left(1+\theta\exp\left(x_i^{'}\beta\right)\right) + \ln\Gamma\left(y_i+\frac{1}{\theta}\right) - \ln\Gamma\left(y_i+1\right) - \ln\Gamma\left(\frac{1}{\theta}\right)$$

Similar to approach in PR model, IRLS algorithm can be used to obtain the solution of this likelihood function and the solution is described as

$$\hat{\beta}_{NBR} = \left(X^{'}\hat{W}_{IRLS-NBR}X\right)^{-1} X^{'}\hat{W}_{IRLS-NBR}\hat{z}$$

where $\hat{W}_{IRLS-NBR}$ and $\hat{z}$ are obtained via IRLS algorithm. In order to improve the conventional NBR model against to multicollinearity problem, NBRR and NBLR model have been propose by Månsson [7] and Månsson [9], respectively. NBRR model is defined as follows:

$$\hat{\beta}_{NBRR} = \left(X^{'}\hat{W}_{IRLS-NBR}X + kI\right)^{-1} X^{'}\hat{W}_{IRLS-NBR}X\hat{\beta}_{NBR},\ k \geq 0$$

NBLR model is also given as

$$\hat{\beta}_{NBLR} = \left(X'\hat{W}_{IRLS-NBR}X + I\right)^{-1}\left(X'\hat{W}_{IRLS-NBR}X + dI\right)\hat{\beta}_{NBR},\ 0 < d < 1$$

Although NBR model is actually effective on dealing with overdispersion problem, some alternatives have been also proposed to improve NBR model's performance. By carrying similar process like in poisson models, zero-inflated negative binomial regression and negative binomial Hurdle regression models are defined as follows:

$$P_{ZNBR}(y_i) = \left\{\begin{matrix} \alpha_i + (1-\alpha_i)\left(\frac{1}{1+\theta\mu_i}\right)^{1/\theta},\ y_i = 0 \\ (1-\alpha_i)\frac{\Gamma\left(y_i+\frac{1}{\theta}\right)}{\Gamma(y_i+1)\Gamma\left(\frac{1}{\theta}\right)}\left(\frac{1}{1+\theta\mu_i}\right)^{1/\theta}\left(\frac{\theta\mu_i}{1+\theta\mu_i}\right)^{y_i},\ y_i \geq 1 \end{matrix}\right\}$$

$$P_{NBHR}(y_i) = \left\{\begin{matrix} \alpha_i,\ y_i = 0 \\ (1-\alpha_i)\frac{\Gamma\left(y_i+\frac{1}{\theta}\right)^{\alpha_i}}{\left[1-\left(\frac{1}{1+\theta\mu_i}\right)^{1/\theta}\right]\Gamma(y_i+1)\Gamma\left(\frac{1}{\theta}\right)}\left(\frac{1}{1+\theta\mu_i}\right)^{1/\theta}\left(\frac{\theta\mu_i}{1+\theta\mu_i}\right)^{y_i},\ y_i \geq 1 \end{matrix}\right\}$$

where α corresponds the probability of zero counts.

3. Selection Ridge and Liu Parameters

The selection of ridge and Liu biasing parameters significantly affect on the performance of PRR, NBRR, PLR and NBLR models. Although there have been extensive literature on the determination of these parameters, there is no consensus among these studies. Various different methods have been proposed for each model. For the models based on ridge estimators, the most well-known method is defined by Hoerl and Kennard as follows:

$$k = \left(\frac{\hat{\sigma}^2}{\hat{\alpha}^2_{\max}}\right)$$

where $\hat{\alpha} = \gamma\hat{\beta}_{ML}$ and γ corresponds the eigenvector of $X'\hat{W}X$. $\hat{W}$ matrix is obtained via each model, separately. Some alternative estimators have been proposed in order to avoiding underestimating of optimal k value via above method. The following estimator has been suggested by Månsson [6] and Månsson [7] for PRR and NBRR models and considered in this study:

$$k_1 = \max\left(\frac{1}{m_j}\right)$$

where $m_j = \sqrt{\hat{\sigma}^2/\hat{\alpha}^2_j}$.

Similarly, Månsson et al. [8] and Månsson [9] suggested the following selection method on the selection of Liu biasing parameters for PLR and NBLR models:

$$d_1 = \max\left(0, \min\left(\frac{\hat{\alpha}_j^2 - 1}{\left(1/\hat{\lambda}_j\right) + \hat{\alpha}_j^2}\right)\right)$$

where $\hat{\lambda}_j$ is the jth eigenvalue of $X^{'}\hat{W}X$.

4. Performance Evaluation

In order to investigate the performance of regression models, a real data set has been used. The data set has been obtained from Deb and Trivedi's study [21] and is available in [25]. This data set includes 4406 individuals and 12 attributes. The dependent variable is the number of pysician office visits. The number of days of hospital stays, the health status, age, gender, marital status, education, family income, private insurance status, the number of chronic diseases, job status and disability status are considered as the explanatory variables in this study. There is no missing values in data set. The existence of outliers has been examined by using the range of ±3 standard deviation of eah variable. To investigate the existence of outliers, we considered both z-score and individually ±3 standard deviation approach with some graphical methods including box-plot. Mahalanobis distance was used to investigate the multivariate outliers but there have been no critical outliers in terms of this measurement potentially affecting the performance. At the end of univariate outlier deletion process, the remain 4182 individuals are used in experiments.

The data set has been splitted as train and testing data sets with the ratios %70 and %30, respectively. Regression models have been fitted via train data set and tested via testing data set. The distribution of frequencies for each split is given in Fig. 1. All the experimental study has been carried out via R software [26] and related packages including MASS [28], lmtest [29], pscl [30] and countreg [30].

4.1. **Training Regression Models.** In the training phase, six models including PR, ZPR, PHR, NBR, ZNBR and NBHR have been fitted on the training data set. Akaike information criterion (AIC) and log likelihood values have been calculated to determine the best fit. The obtained results has been given in Table 1. According to results in Table 1, NBR, ZNBR and NBHR models which provide the minimum AIC and LL values, have been found as the best models on the training data set. Although the performances of these models are similar to each other, the best one among them is the NBHR model. The worst model for our training data set is the

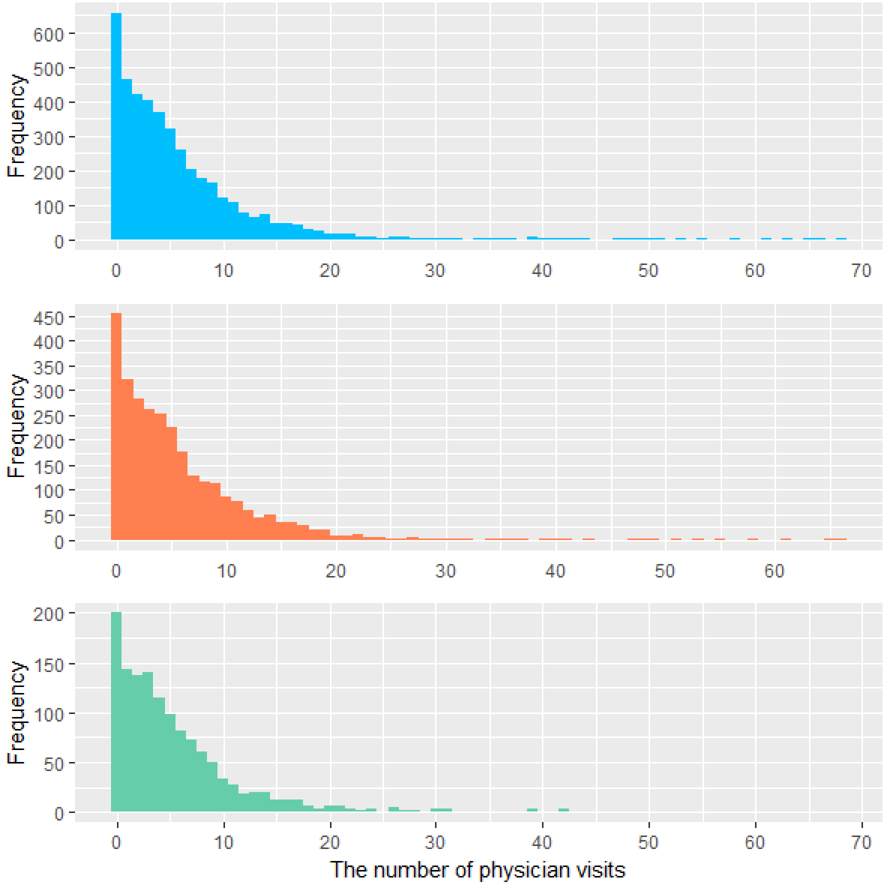

Figure 1. The distribution of frequencies in each data split

classical PR model.

Table 1. AIC and LL values of regression models for training data set

Model	AIC	LL
PR	23239.538	-11606.769
NBR	16091.514	-8031.757
ZPR	20968.239	-10458.119
ZNBR	16003.561	-7974.781
PHR	20968.497	-10458.249
NBHR	15993.192	-7969.596

A log likelihood ratio test has been carried out in nested models to decide whether an overdispersion parameters in negative binomial regression model is necessary. Models based on negative binomial regression and poisson regression have been examined among themselves, separately. The results are given in Table 2. Based on Table 2, it can be said that all comparisons of nested models have been as significant. This result means that models based on negative binomial regression can be seen as more suitable at the point of dealing with overdispersion. Besides, Vuong test [27] is applied training results for non-nested models to determine the model's performance of dealing with excessive zeros. A positive value of test statistic in Vuong test means that the first model is more reasonable than the second one. The test statistics and significance values are given in Table 2. According to Vuong test's results, PHR and ZPR are more preferable as statistically than PR. Similarly, ZNBR and NBHR are found as more suitable than classical NBR model. Only the differences between ZNBR&NBHR and ZPR&PHR are not statistically significant.

Table 2. LRT and Vuong test results based on pair comparisons on training data set

LRT	NBR-PR		ZNBR-ZPR		NBHR-PHR	
Value	7150	(p<0.001)	4966.7	(p<0.001)	4977.3	(p<0.001)
Vuong test	PR-PHR	ZNBR-NBHR	PR-ZPR	NBR-ZNBR	ZPR-PHR	NBR-NBHR
z statistic	-14.183	-1.037	-14.186	-5.008	0.745	-5.352
p	(p<0.001)	(p=0.1499)	(p<0.001)	(p<0.001)	(p=0.2508)	(p<0.001)

As well as LRT and Vuong tests, rootogram graphs have been obtained to examine the model fits as visually. In rootogram graphs, the closeness of the bars to the x axis, is proportional to the goodness of fit. These graphs are given in Figure 2. It can be said that models based on negative binomial regression are seen more better than the models related with poisson regression. Among these models, ZNBR and NBHR are seen more suitable than PR. Although the basic NBR model can be a convincing alternative in the existence of excessive zeros and overdispersion, it's zero inflated and Hurdle variants is more powerful for an effective estimation.

4.2. **Testing Regression Models.** In the training phase, all regression models have been fitted to the training data and some performance measurements like AIC and LL have been given to compare six models except the ones based on ridge and Liu estimators. All outweights vectors have been obtained for all models. By using these outweights, regression models including PRR, PLR, NBRR and NBLR have been tested on the unseen (i.e. testing) data set. The mean squared error (MSE), it's square root (RMSE) and the mean absolute error (MAE) are used as performance criteria. Unstandardized residuals have been considered in the calculation of these measures. The testing results are given in Table 3. Based on Table 3, it is shown that ZNBR, PHR and ZPR models provide better results than their competitors. These models give the smaller MSE, RMSE and MAE values. Although the performances of ZNBR, PHR and ZPR models are similar each other, the best fit is obtained with PHR model. Models based on ridge and Liu estimators are slightly poorer than the rest of models. Overall, the NBLR model is found as

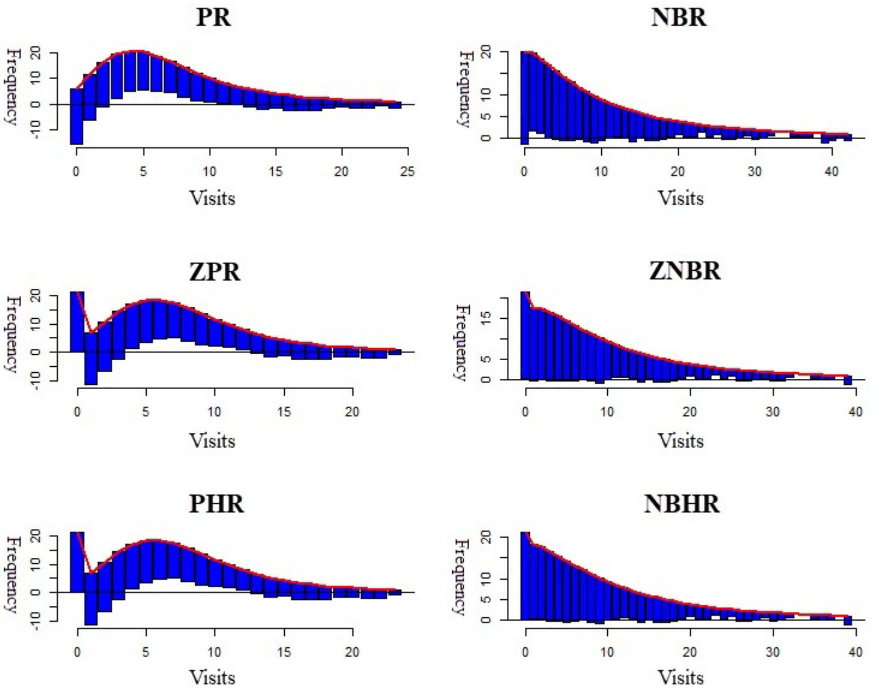

FIGURE 2. The rootogram graphs of each regression model

the worst model for our data set.

Table 3. The comparison of regression models on the testing data set

Model	MSE	RMSE	MAE
PR	39.8572	6.31325	3.9699
NBR	40.7568	6.38410	3.9847
ZPR	39.6489	6.29673	3.9449
ZNBR	39.8083	6.30938	3.9318
PHR	39.6447	6.29640	3.9450
NBHR	40.0699	6.33007	3.9500
PRR	40.8191	6.38898	3.9842
PLR	40.8198	6.38903	3.9844
NBRR	41.0495	6.40698	3.9919
NBLR	41.0499	6.4070	3.9920

4.3. **Determination of the Best Model and Interpretation.** When the results in training and testing phases are examined together with the rootogram graphs, ZNBR model can be seen as the best model for the data set in this study. In the training phase, NBR, ZNBR and NBHR models have similar performances according to the AIC and LL values. NBR, ZNBR and NBHR models show the best fit as visually in rootogram graphs. Finally, ZPR, ZNBR and PHR models are found better than their competitors based on testing performance. Besides, LRT tests is found significant between ZPR and ZNBR models. By virtue of these comments, it can be said that the most suitable model for the data set is ZNBR. The final fit and coefficients have been obtained using ZNBR model. All results of final modelling are given in Table 4. The number of days of hospital stays, the health status, education, age, gender, marital status, family income, private insurance status, the number of chronic diseases, job status and disability status

Table 4. The fitting results of ZNBR model

	Count Part			Zero Part		
Variable	β	exp (β)	Se	β	exp (β)	Se
Intercept	1.593*	4.920	0.21580	4.574*	96.9350	1.45082
#Days of hospital stays	0.280*	1.324	0.02995	-0.533	0.587	0.34171
Health status (perfect)	-0.276*	0.759	0.06445	0.406	1.500	0.28061
Health status (bad)	0.264*	1.303	0.04949	-0.505	0.603	0.58746
#Chronic diseases	0.137*	1.147	0.01308	-1.127*	0.324	0.16908
Disability status (yes)	0.117*	1.124	0.04180	0.039	1.040	0.34310
Age (/10 years)	-0.057*	0.945	0.02730	-0.600*	0.549	0.18999
Gender (male)	-0.026	0.974	0.03552	1.044*	2.840	0.24176
Marital status (married)	-0.111*	0.895	0.03699	-0.747*	0.474	0.25823
Education	0.020*	1.020	0.00471	-0.097*	0.908	0.03041
Family income (x10000$)	0.007	1.007	0.00882	-0.022	0.978	0.06752
Job status (yes)	-0.022	0.978	0.05314	-0.334	0.716	0.37298
Private insurance (yes)	0.137*	1.147	0.04292	-0.172*	0.310	0.22918

Coefficients for each part of zero-inflated negative binomial regression have been given separately. Count part corresponds the modelling results on the individuals who are visited by the pyhsician more than zero. Similarly, zero part results belong the the people who are never visited by the pyhsician. As mentioned before, zero-inflated models present a mixture of different distributions who are able to model count and zero dependent values in a separate process.

According to the count part results in Table 4, the significant variables are #days of hospital stays, health status, #chronic diseases, disability status, age, marital status, education and private insurance ($p < 0.05$). The physicians tend to visit about %24 more to people whose perception of health is perfect than the ones having normal perception. The people having some disabilities in their normal life are more likely to be visited (approximately %15). Similarly, private insurance status is effective on the number of visits. The existence of private insurance provides more visits (%14.7) than normal status. Married people are approximately %10 less likely to be visited than single people. On the other hand, gender, family income and job status are not statistically significant on the number of physician visits ($p < 0.05$).

When observed the results in zero part, it can be seen that #chronic diseases, age, gender, marital status, education and private insurance are statistically significant on the modelling the zero counts ($p < 0.05$). The results show that people with chronic diseases are less likely (approximately %67) not to be visited by the physicians. This results is reasonable because of the risk of those people. The physicians tend not to visit men 2.84 times less likely than women. The comments is supportive with count part when marital status, education and private insurance variable are examined. The more likely situation in count part corresponds the less likely one in zero part. While an occurrance for a particulary variable in count part can increase the possibility of visits, it can decrease for the zero-part or vice versa. Compared with count part, there is more insignificant variables in zero part. #days of hospital stays, health status, disability status, family income and job status are not the effective factors ($p < 0.05$).

5. Conclusions and Discussions

In this study, we have used poisson and negative binomial regression models and their variants based on zero-inflated, Hurdle, ridge and Liu approaches to model a medical data. Zero-inflated, Hurdle and biased (i.e. ridge and Liu) models have been considered in order to get better solution in the presence of excessive zeros and overdispersion. Because of the prevalence of these problems in real word application, classical regression models can be a remedy. As a result of comprehensive experimental study, it can be said that zero-inflated and Hurdle models may provide better results than their competitors. The researchers facing excessive zeros and overdispersion problems may consider these models as powerfull alternatives to get useful insight about their application. Moreover, poisson and negative binomial regression models based on ridge and Liu estimators should be taken into consideration in the presence of multicollinearity in addition to excessive zeros and overdispersion problems.

Although the mentioned models can provide useful insights for modeling count data, it should be noted that the performance comparison is valid for data sets having similar underlying structure. In the future works, we will carried out a comprehensive simulation study to achieve better generalization performance. Different types of data sets belonging various domains, larger dimensions and more detailed parameter selection process can be seen as the limitations of this study.

Author Contribution Statements Gizem YILDIRIM: Writing- Original draft preparation, Investigation, Data curation. Selahattin KAÇIRANLAR: Conceptualization, Methodology, Supervision, Validation, Writing-Reviewing and Editing. Hasan YILDIRIM: Software, Formal analysis, Resources, Visualization.

Declaration of Competing Interests The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Dobson, A. J., Barnett, A. G., An Introduction to Generalized Linear Models, Chapman and Hall/CRC, 2008. https://doi.org/10.1201/9781315182780

[2] Chatterjee, S., Hadi, A. S., Regression Analysis by Example, John Wiley & Sons, 2015. https://doi.org/10.1002/0470055464

[3] Cameron, A. C., Trivedi, P. K., Regression Analysis of Count Data, 6th edn., Cambridge University Press, New York, 2007. http://dx.doi.org/10.1017/CBO9780511814365

[4] Hoerl, E., Kennard, R. W., Ridge regression: biased estimation for nonorthogonal problems, *Technometrics,* 12 (1970), 55–67. https://doi.org/10.1080/00401706.1970.10488634

[5] Liu, K., A new class of biased estimate in linear regression, *Commun. Stat. Theory Methods,* 22 (1993), 393-402. https://doi.org/10.1080/03610929308831027

[6] Månsson, K., Shukur, G., A Poisson ridge regression estimator. *Econ. Model.,* 28 (2011), 1475-1481. https://doi.org/10.1016/j.econmod.2011.02.030

[7] Månsson, K., On ridge estimators for the negative binomial regression model, *Econ. Model.,* 29 (2012), 178-184. https://doi.org/10.1016/j.econmod.2011.09.009

[8] Månsson, K., Kibria, B. M. G., Sjolander, P., Shukur, G., Improved Liu estimators for the poisson regression model. *Int. J. Stat. Probab.,* 1 (2012), 2-6. https://doi.org/10.5539/ijsp.v1n1p2

[9] Månsson, K., Developing a Liu estimator for the negative binomial regression model: method and application. *J. Stat. Comput. Simul.,* 83 (2013), 1773-1780. https://doi.org/10.1080/00949655.2012.673127

[10] Wang, W., Famoye, F., Modeling household fertility decisions with generalized Poisson regression, *Journal of Population Economics*, 10(3) (1997), 273-283. https://doi.org/10.1007/s001480050043

[11] Famoye, F., Singh, K. P., Zero-inflated generalized Poisson regression model with an application to domestic violence data, *Journal of Data Science,* 4(1) (2006), 117-130. https://doi.org/10.6339/JDS.2006.04(1).257

[12] Bandyopadhyay, D., DeSantis, S. M., Korte, J. E., Brady, K. T., Some considerations for excess zeroes in substance abuse research, *The American Journal of Drug and Alcohol Abuse*, 37(5) (2011), 376-382. https://doi.org/10.3109/00952990.2011.568080

[13] Buu, A., Johnson, N. J., Li, R., Tan, X., New variable selection methods for zero-inflated count data with applications to the substance abuse field, *Statistics in Medicine*, 30(18) (2011), 2326-2340. https://doi.org/10.1002/sim.4268

[14] Mouatassim, Y., Ezzahid, E. H., Poisson regression and zero-inflated Poisson regression: application to private health insurance data, *European Actuarial Journal,* 2(2) (2012), 187-204. https://doi.org/10.1007/s13385-012-0056-2

[15] Xie, H., Tao, J., McHugo, G. J., Drake, R. E., Comparing statistical methods for analyzing skewed longitudinal count data with many zeros: An example of smoking cessation, *Journal of Substance Abuse Treatment,* 45(1) (2013), 99-108. https://doi.org/10.1016/j.jsat.2013.01.005

[16] Liyanage, T., Ninomiya, T., Jha, V., Neal, B., Patrice, H. M., Okpechi, I., Zhao, M-H., Lv, J., Garg, A. X., Knight, J., Rodgers, A., Gallagher, M., Kotwal, S., Cass, A., Perkovic, V., Worldwide access to treatment for end-stage kidney disease: a systematic review, *The Lancet,* 385(9981) (2015), 1975-1982. https://doi.org/10.1016/s0140-6736(14)61601-9

[17] Martinez, F. J., Calverley, P. M., Goehring, U. M., Brose, M., Fabbri, L. M., Rabe, K. F., Effect of roflumilast on exacerbations in patients with severe chronic obstructive pulmonary

disease uncontrolled by combination therapy (REACT): a multicentre randomised controlled trial, *The Lancet,* 385(9971) (2015), 857-866. https://doi.org/10.1183/13993003.00158-2017

[18] Oliveira, M., Einbeck, J., Higueras, M., Ainsbury, E., Puig, P., Rothkamm, K., Zero-inflated regression models for radiation-induced chromosome aberration data: A comparative study, *Biometrical Journal,* 58(2) (2016), 259-279. https://doi.org/10.1002/bimj.201400233

[19] Tang, Y., Liu, W., Xu, A., Statistical inference for zero-and-one-inflated poisson models, *Statistical Theory and Related Fields,* 1(2) (2017), 216-226. https://doi.org/10.1002/bimj.201400233

[20] Chai, T., Xiong, D., Weng, J., A zero-inflated negative binomial regression model to evaluate ship sinking accident mortalities, *Transportation Research Record*, 2672(11) (2018), 65–72. https://doi.org/10.1177

[21] Deb, P., Trivedi, P. K., Demand for medical care by the elderly: a finite mixture approach, *Journal of Applied Econometrics*, 12(3) (1997), 313-336. http://www.jstor.org/stable/2285252?origin=JSTOR-pdf

[22] Garthwaite, P. H., Jolliffe, I. T., Jones, B., Statistical Inference, Oxford University Press, Oxford, 2002. https://doi.org/10.1017/S0025557200173425

[23] Hilbe, J. M., Negative Binomial Regression, Cambridge University Press, Cambridge, 2011. https://doi.org/10.1017/CBO9780511973420

[24] Lambert, D., Zero-inflated Poisson regression with an application to defects in manufacturing, *Technometrics,* 34 (1992), 1-14. https://doi.org/10.2307/1269547

[25] https://www.jstatsoft.org/article/view/v016i09

[26] Core Team, R., R: A language and environment for statistical computing, Vienna: Austria: R foundation for Statistical Computing, (2016). http://www.R-project.org/

[27] Vuong, Q. H., Likelihood ratio tests for model selection and nonnested hypotheses, *Econometrica,* 57(2) (1989), 30-33. https://doi.org/10.2307/1912557

[28] Venables, W. N., Ripley, B. D., Modern Applied Statistics with S, Fourth Edition, Springer, New York, 2002. https://www.stats.ox.ac.uk/pub/MASS4/.

[29] Zeileis, A., Hothorn, T., Diagnostic checking in regression relationships, *R News*, 2(3) (2002) 7-10. https://CRAN.R-project.org/doc/Rnews/.

[30] Zeileis, A., Kleiber C, Jackman, S., Regression models for count data in R, *Journal of Statistical Software*, 27(8) (2008), 1-25. http://www.jstatsoft.org/v27/i08/.