International Journal of Statistics and Systems
ISSN 0973-2675 Volume 12, Number 2 (2017), pp. 265-276

http://www.ripublication.com

A Comparative Study of Zero-inflated, Hurdle Models with Artificial Neural Network in Claim Count Modeling

K. M. Sakthivel[1] and C.S.Rajitha[2]

Department of Statistics, Bharathiar University, Coimbatore, Tamilnadu- 641046, India

Abstract

Modeling of claim frequency is a vital factor in the non-life insurance industry. The claim count data in non-life insurance may not follow the traditional regression count data models with the use of Poisson or negative binomial distribution in numerous circumstances due to excessive number of zeros in the real data set. If the excessive number of zeros is not considered with sufficient weight, it will lead to information shortage to get a accurate rate making for the non-life insurance portfolio. In this paper we compared different claim count models such as zero-inflated Poisson (ZIP) regression model, hurdle model with back propagation neural network (BPNN) for modeling the count data which has excessive number of zeros. We shown from our empirical study that BPNN outperforms the conventionally used models and provided better fit to claim count data in terms of mean squared error (MSE).

Keywords: Artificial Neural Network; Back Propagation Algorithm; Claim frequency; Mean Squared Error

1. INTRODUCTION

In non-life insurance, potential profits are determined by a sequence of income payments called premiums and outgoing payments called claims. Therefore estimating the accurate premium expenses is the most vital task in the general insurance. To model the distribution for pure premiums, the traditional two part model method is usually used by disintegrating the total payments into number of claims or

claim frequency or claim counts and the amount of claim or claim severity (Klugman et.al, 2008). In this paper the focal point of concentration is on the claim frequency which is the most essential part of setting the premium. Therefore modeling the claim count distribution is an important and difficult task in general insurance practice. The traditional statistical methods assumes that the number of claims follow the Poisson distribution (Antonio and Valdez, 2012; Renshaw, 1994; Cameron and Trivedi, 1986). A commonly used method preferred for modeling claim frequency or claim count is the regression model, where individual features are taken into account by incorporating a regression constituent. Denuit, et. al (2007) used a particular type of generalized linear model (GLM) called Poisson regression model for modeling claim frequency distribution. But most of the time insurance data holds excessive number of zeros, since many policies did not have claim for a specific time period. For instance there is a possibility that policyholders doesn't report the small claims for getting no claim bonus and deductibles for reducing the premium of forthcoming year (Yip and Yau, 2005), due to this reason Poisson regression models may not give adequate result in this circumstance. Further the most important equidispersion property of Poisson distribution is violated. Then for handling the over dispersed count data, negative binomial model was formulated as an alternative to Poisson model but this model also does not provide exact prediction in certain situations.

To overcome the limitation of these regression models, Lambert (1992) presented a more realistic way for modeling count data which has large counts and zeros, called zero-inflated Poisson (ZIP) regression models. Numerous alterations of the Poisson regression model have been presented by Greene (1994) as an extension to Lambert's ZIP regression model. A number of parametric zero-inflated count distributions have been presented by Yip and Yao (2005) to provide accommodation to the surplus zeros to insurance claim count data. Flynn (2009) made a comparative study of zero-inflated models with conventional GLM frame work having negative binomial and Poisson distribution choice. Instead of taking the excess number of zeros in one part and a standard count distribution such as regular Poisson or negative binomial distribution in the another part, hurdle model take account of all zeros in the right censored part and all positive counts in the left truncated-at-zero part. The hurdle models for count data has been first discoursed by Mullahy (1986). Heilbron (1994) called this model as a two part model. For handling the over dispersion and under dispersion present in the data, Gurmu (1998) initiated the generalized hurdle model. Saffari, et.al (2012) recommended censored negative binomial hurdle model specification to model the count data with excessive number of zeros which successfully deal with over dispersion problem. The difference between zero-inflated models and hurdle models have been re-examined by Loeys, et.al, (2012) in their tutorial. Baetschmann & Winkelmann (2014) derived a new approach for the modelling of zero-inflated count data, called dynamic hurdle model and giving new justifications to excessive zeros and discussed its properties. They assumed that the counts are produced from the non-stationary stochastic process.

But all of these models failed to capture the latent dynamics present in the data despite of giving better fit to data with excess number of zeros. Therefore Artificial

Neural Networks (ANN) can be used as an alternative since it captures such phenomena and provides accurate prediction of future claim counts as the sample size increases. Earlier Sarle (1994) established the interconnection among ANN and traditional statistical modeling techniques such as GLM, cluster analysis, maximum redundancy analysis etc. Also they made clear that when evaluating the data, statistics and neural networks are not contenting methods and there is substantial intersection between both fields. Brocket, et. al (1994) introduced an artificial neural network with back propagation (BP) algorithm for predicting the insurer insolvency and highlighted the effectiveness of this method compared to discriminant analysis. Warner and Misra (1996) examined superiority of ANN over regression models and also discussed the difficulties of implementing the ANN. According to Zhang, Patuwo, & Hu, (1998) performance of ANN is comparatively better and adaptable than other forecasting methods. Dalkilic, et. al (2009) pointed out the reasons for using neural network (NN) approach with fuzzy rules instead of least square method, when there is at least one outlier exist in the claim payments. And he developed an algorithm using adaptive network for the determination of regression parameters. Bahia (2013) showed that ANN provides results for estimating and forecasting insurance premium revenue. Bapat, et. al (2010) formulated an effective ANN with BP algorithm for predicting the motor insurance claims for forthcoming years based on the past years information. Recently Soni, et.al (2015) suggested an ANN with resilient BP algorithm for assessing credit applications and demonstrated that this model assisting loan determinations in a well manner in commercial banks. In 2016 Yunos, et. al suggested that neural network with BP algorithm can be used as a technique for handling the insurance data which has vast information, dubiousness and incomplete information. Recently Sakthivel and Rajitha (2017) developed a procedure for predicting the future claim frequency of an insurance portfolio in general insurance using ANN.

In this paper, zero-inflated insurance claim count data is modeled using artificial neural network, ZIP regression and hurdle models. And we compared the efficiency of these models using mean square error (MSE). The arrangement of the paper is as follows: section 2 provides detailed description about ZIP regression models and section 3 highlights about hurdle Poisson regression models. Section 4 provides details of neural network computation and different types of neural networks. Section 5 provides empirical study on claim count data. Section 6 &7 provides results and conclusions about this study.

2. ZERO-INFLATED MODELS

The claim count data in general insurance do not follow the classical Poisson regression model because it exhibits excess number of zero counts called zero inflation in most of the cases. And the Poisson regression model is not appropriate in the case of zero-inflated data due to the destruction of the equidispersion (i.e., mean = variance) property. To overcome this difficulty, Lambert (1992) introduced an

alternative model called zero-inflated Poisson regression model. The design of this model is of two fold. First one is the modeling of zero counts by admitting the excess zero ratio π and the proportion $(1-\pi)(e^{-\lambda})$ for zeros coming from the Poisson distribution and the next model for positive counts using a zero-truncated Poisson model. The specification of the ZIP regression model is as follows

$$P(Y=y/\lambda,\pi)=\begin{cases}\pi+(1-\pi)e^{-\lambda} & when \;\; y=0\\ (1-\pi)\dfrac{e^{-\lambda}\lambda^{y}}{y!} & when \;\; y>0\end{cases}$$

λ is the mean of the Poisson distribution and π is the probability value of the extra zero counts. The first two moments of the ZIP regression model are

$$E(Y)=(1-\pi)\lambda$$

$$V(Y)=\lambda(1-\pi)(1+\lambda\pi)$$

3. HURDLE MODELS

Mullahy (1986) developed hurdle models to take care of the excess zero counts when the data generating process give more number of zeros than expected by the distributional assumptions of standard count distributions. Hurdle model evaluated zero and non-zero counts independently for modeling the zero-inflated counts and all zero counts are considered as true zeros. The hurdle models begin by means of the binomial practice, which find out whether the count (response) variable obtains the value zero or a positive value. The major advantage of using a hurdle model is that it allows the statistical procedure for an organized distinction to manage observations under the hurdle with a zero count model and over the hurdle with a zero-truncated count model. Therefore the hurdle models are also called two part models (Heilbron, 1994). Usually the second part of the model is in the zero truncated structure of traditional standard distributions like Poisson or negative binomial. Therefore in the literature commonly used types of specifications for hurdle models are Poisson hurdle specification and negative binomial hurdle specification. In this paper, we considered the hurdle Poisson model specification. The hurdle Poisson regression model with count variable y has the distribution

$$P(Y=y/\lambda,\pi_0)=\begin{cases}\pi_0 & ; \quad y=0\\ \dfrac{(1-\pi_0)e^{-\lambda}\lambda^{y}}{(1-e^{-\lambda})y!} & ; \quad y>0\end{cases}$$

If $y>0$ means the hurdle is crossed then the conditional distribution of the non-zero values is managed by a zero truncated count model. λ is the mean of the Poisson distribution and π_0 is the probability value of the zero counts. For estimating the parameter values, maximum likelihood method (MLE) is used. This model is nothing

but a reparameterization of the ZIP regression model. Although for both models the parameters are modeled in the regression framework, hurdle Poisson regression model is not as same as the ZIP regression model.

4. NEURAL NETWORK

In the modeling of insurance claim count data the neural network has an excellent potential because it perform linear and nonlinear mapping without any preliminary information exists in the data. Therefore ANN provides flexible, consistent and reliable appraisals compared to other statistical methods used for modeling the claim count data. The key difference between ANN and traditional methods is that traditional methods consider mathematical logic for modeling whereas NN learns inductively or empirically thereby training the network. Also it learns the logic underlying between the input and the observed output by the training process and develops a mathematical linear / nonlinear network. NN does not require a priori model for clarifying the correlation between input variables and target variables, but the fundamental statistical models have need of this information. And also it surely gives permission to the network to adapt to new situations due to its dynamic learning process. Unlike other methods, the entire network will be stable even during uncertain/ extreme behavior of data. Studies relating to the network show that predictions are rather insensible to deviations involved in the network pattern (Neti, Schneider, and Young, 1992).

4.1 Network Architecture

ANNs are one of the most popular machines learning method which are able to do classification and prediction tasks in an exact and more reliable manner. According to Simon S. Haykin (2009), ANN form a directed graph by connecting the artificial neurons, the basic processing components of the network. The three basic elements of a neural network are the basic computing elements, known as neurons or nodes, the network architecture which describes the association between computing units or neurons and the training algorithm used to find the weights which modifies the strength of the input for performing a particular task. Architecture of the network refers to the arrangement of the units and the types of connections permitted. In the multilayer feed forward network, units are ordered in a series of layers, this is one of the network type used most often in statistical applications. The movement of information is from lower layers to the higher layers of the network. The weights are usually obtained by optimizing the performance output of the network on a set of training examples with respect to some loss or error function. The standard structure of the network is given in the Figure 1.

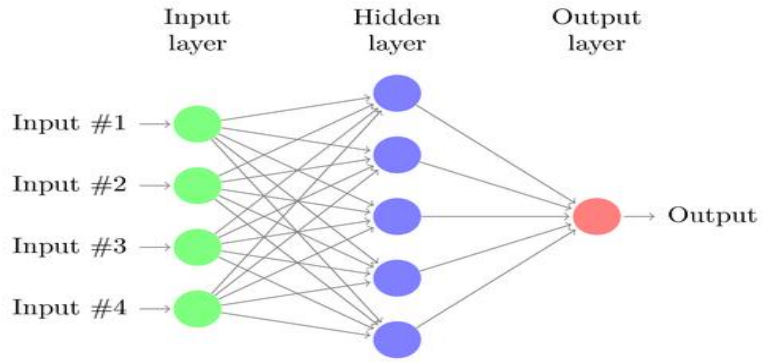


Figure 1

4.2 Multilayer Feed Forward Networks

The multilayer feed forward NNs, which are used most often in statistical applications consist of several layers usually three or four layers of computing units. Network receives information only from computing units in the first layer. Hence the first layer serves as the input layer. The model solution is produced by the last layer known as the output layer. Intermediate layers are the hidden layers and they are critical for NN to identify the complex pattern in the data. The units in the hidden layers are called hidden units since they are not visible to the user in the sense that input units and output units are visible. In a theoretical manner, multilayered feed forward NNs are universal approximators, and with respect to its inherent nature, it has a tremendous capacity of constructing any nonlinear mapping to any extent of accuracy (Hornik et. al, 1989). They do not need a priori model to be assumed or a priori assumptions to be made on the properties of data (C. M. Bishop, 1995). They have been widely employed for modeling, prediction, classification, optimization, and control purposes (M. Azlan Hussain, 1999; J. G. De Gooijer and R. J. Hyndman, 2006; B. K. Bose, 2007).

5. DATA ANALYSIS

In this paper, the modeling of claim count is based on the simulated car insurance data set taken from the package *InsuranceData* in R software. The data set consists of a claim file with 1,20,000 records (total records for three years) with 40,000 policies each year. The outcome of interest is the number of claims based on three categorical (independent) variables, driver's age category, vehicle value and the period. At first we decided to model the claim count data using three different methods, ZIP regression model, hurdle Poisson regression model and BPNN. Figure 2 represents the frequency plot of the count variable from the simulated car insurance data set and it shows that large proportion of zeros exist in the data. Table 1 shows the claim frequencies of the data and observed that the data contains 85.73% of the values are zero.

Table 1: Counts of claim frequencies

Claim Frequency	Count	Percentage	Claim Frequency	Count	Percentage	Claim Frequency	Count	Percentage
0	102870	85.725	12	19	0.016	25	4	0.0033
1	11872	9.8933	13	20	0.017	26	1	0.00083
2	2995	2.496	14	8	0.007	27	2	0.0017
3	1029	0.8575	15	6	0.005	29	1	0.00083
4	457	0.38	16	8	0.007	30	1	0.00083
5	260	0.2167	17	6	0.005	32	1	0.00083
6	140	0.1167	18	4	0.0033	33	1	0.00083
7	96	0.08	19	3	0.0025	36	1	0.00083
8	63	0.053	20	6	0.005	37	1	0.00083
9	51	0.04	21	4	0.0033	38	1	0.00083
10	35	0.03	22	3	0.0025	43	1	0.00083
11	25	0.021	23	5	0.0042			

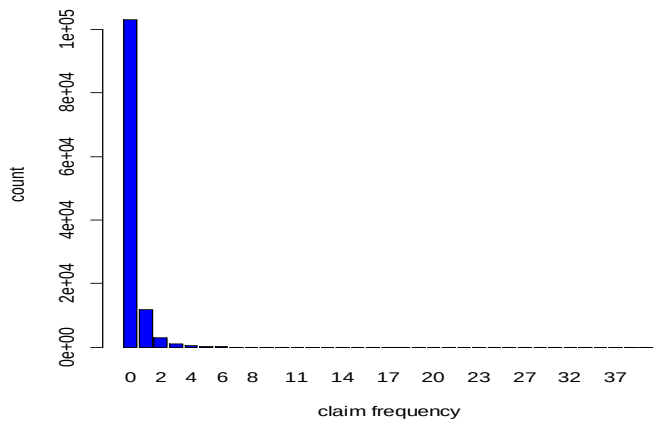


Figure 2. Frequency distribution of the number of claims

For modeling the claim counts needed to select the input variables or rating factors. Table 2 shows the descriptions of the input variables and output variable for modeling the claim count data. Then the data is partition as training set and testing set. There is no rule available so far in the literature for optimum number of observations in training and test set. However Zhang, et.al (1998) recommended some ratios for partitioning the data set, which are 70 : 30, 80 : 20 and 90 : 10. In this paper, we used 80 : 20 ratio for dividing the data set. Data partition is given in the table 3.

Table 2: Type of variables

Input variables	Output variable
1) Driver's age category 2) Vehicle value 3) Period	1) Number of claims

Table 3: Ratio of partition of data

Classification	% of partition of data	Number of observations
Training	80%	96,000
Testing	20%	24,000
Total	100%	120,000

5.1 Estimation of claim count using ZIP regression and hurdle models

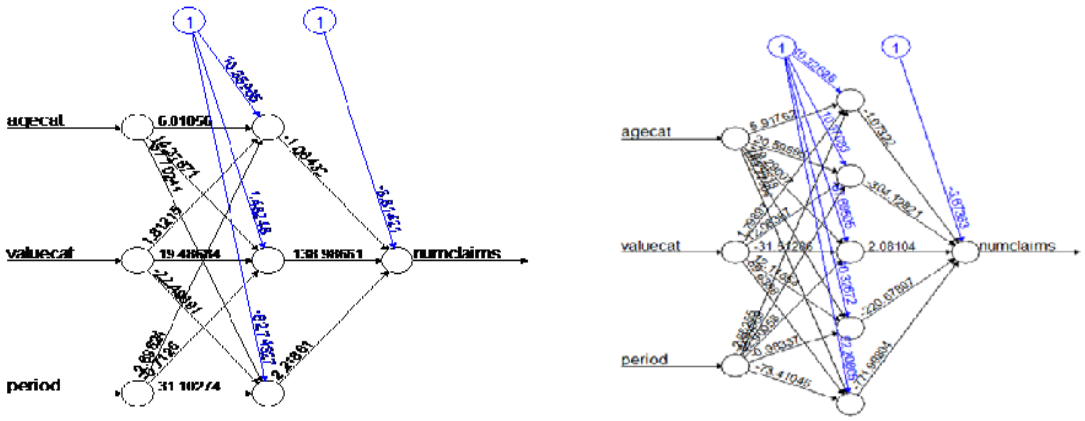
From the above data, we observed around 86% of the data having values zero. Hence for modeling this claim count data, zero-inflated models are more appropriate. Further, the number of claims is considered as the response variable and drivers age category, vehicle value and period (3 years) are considered as the independent variables. And for estimation of claim frequency, we applied ZIP regression and hurdle Poisson regression model. The measure used for efficiency of estimates, we have obtained the MSE for both models.

5.2 Estimation of the claim count using ANN

Back propagation learning algorithm is long been used in the neural network for its reliability of fast convergence. The BPNN encompasses an input layer, an output layer and a hidden layer. Here, the input layer consisted of information from the simulated car insurance data set, they are driver's age category, vehicle value and period and the number of claims is considered as the output layer. By modifying the weights of the BPNN by using a trial and error method, the output can be improved. Here we considered single hidden layer BPNN and double hidden layer BPNN with different number of neurons in each hidden layer and evaluate the accuracy of estimate by using actual claim counts and estimated claim counts interms of MSE. Table 4 represents the network structures used in this study. Figure 3 and Figure 4 shows the network structures with different number of hidden layers and different number of neurons in each hidden layer.

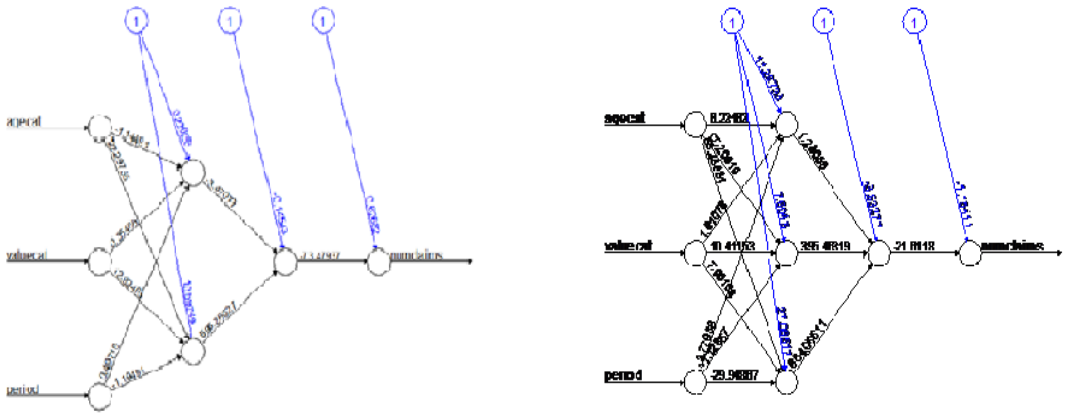
Table 4: Structure of Neural Networks

Two hidden layers	Single hidden layer
3-2-1-1	3-3-1
3-3-1-1	3-5-1



3-3-1 3-5-1

Figure 3 : Structure of single hidden layer neural network



3-2-1-1 3-3-1-1

Figure 4 : Structure of two hidden layer neural network

6. MODEL PERFORMANCE ANALYSIS FOR CLAIM COUNT DATA

Modeling of claim frequency is done by using three methods namely ZIP regression model, hurdle Poisson regression model and ANN for simulated car insurance data. The estimation of claim frequency using ZIP regression and hurdle Poisson regression model is performed using R software. Further for BPNN, four networks have been identified based on our experience and applied feed forward network with back propagation algorithm. Mean square error is calculated for comparing the efficiency of these models. Table 5 shows the MSE of estimates of claim frequencies for ZIP regression and hurdle Poisson regression model and Table 6 shows the MSE of estimates of claim frequency for ANN for different structure of networks. From the Table 5 and Table 6, we concluded that ANN performs better than the conventional ZIP regression and hurdle Poisson regression model for estimation of claim frequency

since MSE of all the four selected structure of ANN is smaller or equal compared to ZIP regression and hurdle Poisson regression model.

Table 5: Performance of ZIP regression and hurdle Poisson regression model

Model	MSE
ZIP	0.73195
hurdle Poisson	0.8478

Table 6: Performance of ANN

Number of hidden layers	Number of Neurons	MSE
One	h = 5	0.73193
One	h = 3	0.73000
Two	h = c(2,1)	0.15797
Two	h = c(3,1)	0.73195

7. CONCLUSION

In this paper, we compared three different models for estimation of the claim count data with excess number of zeros namely ZIP regression model, hurdle Poisson regression model and ANN. We observed that the ANN performs better for estimation of claim frequency compared to the specially made probability models such as ZIP regression and hurdle Poisson regression for zero-inflated claim count data due to its flexibility and adaptive learning properties. Hence the results of ANN can be further improved by modifying the network structure and fine tuning the ratio of training and testing data. Though ANN suffers by black box syndrome, recent research showed that ANN has invincible place as its role in classification and prediction in modern day computation.

REFERENCE

[1] Antonio, K., and Valdez, E.A., 2012, " Statistical Concepts of a Priori and a Posteriori Risk Classification in Insurance". Advances in Statistical Analysis., 96(2), pp. 187-224

[2] Azlan Hussain, M., 1999, "Review of the applications of neural networks in chemical process control – simulation and online implementation," Artificial Intelligence in Engineering., 13, pp. 55–68,.

[3] Bahia, I. S. H., 2013," Using Artificial Neural Network Modeling in Forecasting Revenue: Case Study in National Insurance Company/Iraq," International Journal of Intelligence Science., 3, 136-143.

[4] Bapat, A., and Bapat, P., 2010, "Development and Testing of an Efficient Artificial Neural Network Algorithm and its Effectiveness for Prediction of Insurance

Claims,"International journal of Innovation Management and Technology, 1(3), pp. 286-291.

[5] Baetschmann, G., and. Winkelmann, R., 2014, "A Dynamic Hurdle Model for Zero-Inated Count Data With an Application to Health Care Utilization," Working Paper Series, ISSN 1664-705X (online), Working Paper No. 151.

[6] Bishop, C. M., 1995, Neural Networks for Pattern Recognition, Oxford University, Press, Oxford,.

[7] Bose, B. K., 2007, "Neural network applications in power electronics and motor drives—an introduction and perspective," IEEE Transactions on Industrial Electronics., 54, pp. 14–33

[8] Brockett, P. L., Cooper, W. W., Golden, L. L., and Pitaktong, U., 1994, "A Neural Network Method for Obtaining an Early Warning of Insurer Insolvency," The Journal of Risk and Insurance, 61(3), pp. 402-424.

[9] Cameron, A.C., and Trivedi, P.K., 1986, "Economic models based on count data: Comparisons and applicationsof some estimators and tests," Jounral of Applied Econometrics., 1, pp. 29-53.

[10] Dalkilic T.E., Tank, F., and Kula, K. S., 2009, "Neural networks approach for determining total claim amounts in insurance," Insurance: Mathematics and Economics., 45, pp. 236-241.

[11] Denuit, M., Marechal, X., Pitrebois, S., and Walhin, J- F., 2007,Actuarial Modelling of Claim Counts: RiskClassification, Credibility and Bonus-Malus Systems. John Wiley & Sones Ltd.: West Sussex, England.

[12] De Gooijer, J. G., and Hyndman, R. J., 2006, "25 years of time series forecasting," International Journal of Forecasting., 22(3), pp. 443–473.

[13] Flynn, M., and Francis, L.A., 2009, "More Flexible GLMs: Zero-Inflated Models and Hybrid Models," Casualty Actuarial Society E-Forum., pp. 148-224 .

[14] Greene, W., 1994, "Accounting for Excess Zeros and Sample Selection in Poisson and Negative Binomial Regression Models,"NYU Working Paper, No. EC-94- 10

[15] Greene, W., 2007, "Functional Form and Heterogeneity in Models for Count Data," Foundations and Trends in Econometrics., 1(2), pp. 113-218.

[16] Gurmu, S., 1998, "Generalized hurdle count data regression models,"Economic Letters., 58, pp. 263-268.

[17] Haykin, S. S., 2001, Neural Networks, A comprehensive foundation, Second Edition, Prentice Hall, Pearson Education, Singapore.

[18] Haykin, S. S., 2009, Neural networks and learning machines. Pearson, Upper Saddle River, N.J.

[19] Heilbron, D., 1994, "Zero-altered and other regression models for count data with added zeros," Biometrical Journal., 36, pp. 531–547.

[20] Hornik, K., Stinchcombe, M., and White, H., 1989, "Multilayer feedforward networks are universal approximators," Neural Networks., 2(5), pp. 359–366.

[21] Klugman, S. A., Panjer, H. H., and Willmot, G. E., 2008,Loss Models: From Data to Decisions (Thirded.). John Wiley & Sons, Inc.

[22] Lambert, D., 1992, "Zero-Inflated Poisson Regression, with an Application to Defects in Manufacturing," Technometrics., 34(1), pp. 1-17.

[23] Loeys, T., Moerkerke, B., Smet, O. D., and Buysse, A., 2012, "The analysis of zero-inflated count data: Beyond zero-inflated Poisson regression," British Journal of Mathematical and Statistical Psychology., 65, pp. 163-180.

[24] Mullahy, J., 1986, "Specification and Testing of some modified count data models," Journal of Econometrics., 33, pp. 341-365

[25] Neti, C., Schneider, M. H., and Young, E. D., 1992, "Maximally Fault Tolerant Neural Networks," IEEE Transactions on Neural Networks., 30(1), pp. 12-23,.

[26] Rosenblatt, F., 1958, "The perceptron: A probabilistic model for information storage and organization in the brain,"P sychological Review., 65(6), pp.386-408.

[27] Renshaw, A. E., 1994, "Modelling the claims process in the presence of covariates," ASTIN Bulletin., 24, pp. 265-285.

[28] Sakthivel, K.M., and Rajitha. C.S., 2017, "Artificial Intelligence for Estimation of Future Claim Frequency in Non-Life Insurance," Global Journal of Pure and Applied Mathematics., 13(6), pp. 1701-1710.

[29] Saffari, E.S., Adnan, R., and Greene, W., 2012, "Hurdle negative binomial regression model with right cencored count data," SORT., 36(2), pp.181-194.

[30] Sarle, W. S., 1994, "Neural networks and statistical models," In Proceedings of the 19th annual SAS users group international conference.

[31] Soni, A.K., and Abdullahi, A. U., 2015, "Using Neural Networks for Credit Scoring," International journal of Science, Technology and Management., 4(5), pp. 26-31.

[32] Stergious, C., and Siganos, D., 2007, "Neural Networks," Available: www.Doc.k.ac.uk/and/surprise96/journal /Vol4/cs11/report.html

[33] Warner, B., and Misra, M., 1996, "Understanding neural networks as statistical tools," The American Statistician., 50(4), pp.284–293.

[34] Yip, K.C.H., and Yau, K.K.W., 2005, "On modeling claim frequency data in general insurance with extra zeros,"Insurance: Mathematics and Economics., 36, pp.153-163.

[35] Yang, Yi., Qian, Wei., and Zou, Hui., 2016, "Insurance Premium Prediction via Gradient Tree- Boosted Tweedie Compound Poisson Models," Journal of Business and Economic Statistics., 34(3), pp. 1-45.

[36] Yunos, Z.M., Ali, A., Shamsyuddin, S.M., Ismail, N., and Sallehuddin, R. S., 2016,"Predictive Modelling for Motor Insurance Claims Using Artificial Neural Networks," International Journal of Advances in Soft Computing and its Applications., 8(3), pp.160-172.

[37] Zhang, G., Patuwo, B. E., Hu, M.Y., 1998, "Forecasting with artificial neural networks: The state of the art," International Journal of Forecasting., 14, pp.35–62.

*****