Journal of Contemporary Issues in Business and Government Vol. 29, No. 03, 2023
https://cibgp.com/

P-ISSN: 2204-1990; E-ISSN: 1323-6903
DOI: 10.47750/cibg.2023.29.03.017

Contemporary Issues
ABDC AUSTRALIAN BUSINESS DEANS COUNCIL
in business and government
ISSN: 1323-6903

Frequency of Indian Health Insurance Claims Data using Zero-Inflated Poisson (ZIP) and Zero-Inflated Negative Binomial (ZINB) Regression Models

[1]**Srimannarayana Gajula,** Assistant Professor,
Institute of Insurance and Risk Management, Hyderabad
Email: gajulasriman19@gmail.com

[2] **Lalitha P S,** Assistant Professor, Department of MBA,
Koshys Institute of Management Studies, Bangalore.
Email: drlalithaps@gmail.com

[3] **Kiran Kumar Paidipati,** Assistant Professor,
Area of Decision Sciences, IIM Sirmaur, Himachal Pradesh
Email: kkpaidipati@iimsirmaur.ac.in

Abstract

The study focused on Indian health insurance claims using the ZIP model (Zero-Inflated Poisson) ZINB model (Zero-Inflated Negative Binomial) and Poisson regression analysis used for measuring the zero-Inflated count data. Frequency of claiming insurance modelled the regression analysis with gender priorities and the situations. The analysis consists of two basic situations based on frequency of claims and priorities of gender. To check the models whether fitted or valid using AIC and -2 Log – likelihood measures and Vuong test statistics to compare the fitted models. The ZIP and ZINB regression models suit for above mentioned situations, to compare the (females and males), secondly (females and males with others, transgenders)

Keywords: Indian Health Insurance, Gender Priority, Poisson, frequency Claims, Zero-Inflated Poisson Regression, Zero-Inflated Negative Binomial Regression,

JEL Classification Number: I100, I130, I150, I180

Introduction:

A health insurance policy serves as a financial buffer for the policy bearer in the event of unforeseeable illnesses that incur high medical expenses. A crucial choice for ensuring that people can enjoy the assurance of a safe future is to purchase health insurance. The guiding premise of insurance is that, out of the overall pool of premium contributions made by policyholders, "the fortunate take care of the unfortunates." With 12% to 15 % of annual

P-ISSN: 2204-1990; E-ISSN: 1323-6903
DOI: 10.47750/cibg.2023.29.03.017

growth rate in Indian non-life insurance strengthen and grown as fastest health insurance industry.

In India, health insurance strengthened and become one of the fastest growing industries in our country. The current study emphases on Indian health assurance entitlements data, which provides information on those who have health insurance coverage through various insurance providers. The nature of data is count-based and displays over- (under-)dispersion with extra zeros. Due to the fact that variance and mean are identical in the distribution, Poisson regression models are capable of handling count data. When data deviates from this presumption, the maximum likelihood method's estimation of standard errors may be partial, and statistics tests that are obtained will be erroneous from the models. To overwhelmed the delinquent of over-dispersion and claim counts, the researchers proposed regression models like Zero Inflated (Poisson, Negative Binomial) for health insurance claim data in India.

Review of Literature:

Several researchers applied the regression models using different types of count data in claims with biomedical, crime, traffic, health care and accidents and other services. Below are few research studies on regression models.

Kibria (2006) observed the fitted parameters Poisson (Po), the ZIP (zero-inflated poisson) and ZINB (zero – inflated Negative Binomial) applied on the RoR (run-off-road) crash data and observed that, Po and ZIP well performed better than NB and ZINB models of regression over discrete count data on inflation. According to the study based on zoological data set.

Ozmen and Famoye in (2007) observed the C. caretta hatchlings being dead due to exposure of sun by using Po, NB, ZIP and ZINB. For estimating the parameters used maximum likelihood (ML) method and for goodness of fit also to assessment the score. Further, GP regression model best suitable to count the data compared to other models.

Zulkifli et al. (2011) applied the phenomenon of handling zero-inflation over dispersion by developing ZIP and ZINB models on crime data. The count data inferred ZINB was best model for handling situations in theft insurance than ZIP model using wald test and likelihood ratio.

Ismail and Zamani (2013), elucidated NB and GP is best suited foe over dispersion or under dispersed count data. Then for mean variance relationship ZINB and ZIGP models suits well. NB and GP fitted to Malaysian OD data claim and ZINB, ZIGP formfitting to German health care count data.

Xu L et al. (2015), applied the regression models to micro biome study and did comparative performance on parametric and non-parametric, hurdle and zero inflated models

P-ISSN: 2204-1990; E-ISSN: 1323-6903
DOI: 10.47750/cibg.2023.29.03.017

using simulations. The study discloses hurdle models and zero-inflated models estimates well compared to other models while using AIC, Vonoge test, controlled types, goodness of fit measures for count data.

Yang et al. (2017), associated various regression models includes (Po), (NB), (ZIP), (ZINB) by using real count data stimulated over zero- inflation and over- dispersion of health surveys.

Astan and Kismiantini (2019), studied health insurance data using binary logistic regression model. It is ownership data of IFLS (Indonesian Family Life Survey). The research inferred that by growing the age the probability of insurers also increases tests using logistic regression. There are many factors significant to ownership are education, marital status, job, impatient care and chronic condition.

Research Methodology:

Data and Sample:

The data used of Health Insurance entitlements of Indian the research is mainly secondary data, attained from IRDAI (Insurance Regulatory Development Authority of India). The period considered for frequency and number of claims from health insurance of India was about five years i.e., 2011 to 2015.

For simulation of data and forming parts are done using software SPSS -20. And, for analysing the data the R programming software is been used. For zero – inflated models *zeroinfl ()* function used from the package *pscl*, whereas Poisson model the *glm()* function used from stats package to fit the data.

Analysis Results:

In this study, the researchers discovered several parameters and inferred about comparison of proposed models based on gender. It is appropriately fit to apply and compare parameters of models proposed for the study. The information regarding the parameters of health insurance data includes claim frequency, gender, age groups, exposure and members of family insured data, displayed in the below Table -1

Table-1: Health Insurance Data Variables analysis description:

Parameters	Categorical/Numeric	Description
Claim of Frequency	Numeric	Total claims numbered in every year (insured).
Gender of Insurers	Categorical	0 for male and 1 for female. (1st part) 1 for others and 0 for transgenders. (2nd part)
Age >30	Categorical	1 for <40, others are 0
Age > 40	Categorical	1 for <50, others are 0

P-ISSN: 2204-1990; E-ISSN: 1323-6903
DOI: 10.47750/cibg.2023.29.03.017

Parameters	Categorical/Numeric	Description
Age > 50	Categorical	1 for <60, others are 0
Age > 60	Categorical	1 for <70, others are 0
Age > 70	Categorical	1 for <80, others are 0
Age < 80	Categorical	1 for >90, others are 0
Exposure	Numeric	one-year insured period is, the value either 0 or
Members of Family	Numeric	Members of family insured.

There are two types of variables considered for the research includes Claim frequency data and the insured data. The data consists of gender, age, exposure and family members are comes under independent variables. Whereas claim frequency is a dependent variable illustrated in table-1.

The data classified into values 0 or 1 along with the description showed in table-1. Additionally, the descriptive statistics of frequency of claims for insured variables like numerical variables displayed in table-2 and categorical variables in table-3.

I) Health Insurance Analysis of Claims between Female and Male:

The table -2 expounded the descriptive statistics of numerical variables with claim frequencies and members of family and exposure as the parameters. overall view of gender of health insurance claims among females and males.

The table -3 elucidated the descriptive statistics of categorical variables and gender and groups of age claims insurance. The data alteration from 0 and 1 was illustrated in table-1.

Table-2: Numerical Variables Descriptive statistics:

Parameters	Minimum	Maximum	Modus	Total
Frequency Claim	0	7	0	9703
Family Members	1	5	1	83981

Parameters	Minimum	Maximum	Median	Total
Exposure	0	1	0.457686	22032.86

Table-3: Descriptive statistics showing Categorical Variables:

Variables	Categories	Explanation	Percentage
Gender	1 0	Female Male	39.77% 60.23%

P-ISSN: 2204-1990; E-ISSN: 1323-6903
DOI: 10.47750/cibg.2023.29.03.017

Variables	**Categories**	**Explanation**	**Percentage**
Age < 30	0 1	Others Age ≤ 30	59.62% 40.38%
Age < 40	0 1	Others Age 31 – 40	79.38% 20.62%
Age < 50	0 1	Others Age 41 – 50	81.78% 18.22%
Age < 60	0 1	Others Age 51 – 60	90.31% 9.69%
Age < 70	0 1	Others Age 61 - 70	91.62% 8.38%
Age > 70	0 1	Others Age > 70	96.41% 3.59%

From table-3, the data interprets very clearly that the percentage of females is quite lesser than males. Majority of the insurers claimed at the age between 40-70 age group. The insurers above 70 years of age group is quite lesser compared to other age groups which infers that >70 age group are very less interested in claiming insurance. At the age of 40 to 70 people mostly preferred to claim insurance and the frequency claim is quite more compare to other age groups.

The statistical tools used in the study is regression model using R programming and SPSS software. The estimation of regression model performed firstly with the assistance of regression model Poisson. If the data pertaining more numeral like zeros, with ZIP (Zero Inflated Poisson) and ZINB (Zero Inflated Negative Binomial) the regression analysis accomplished which elucidated in below table-4.

Table 4: Estimation Model for Poisson, ZIP and ZINB regression:

Constraints		**Poisson**	**ZIP**		**ZINB**	
		Model Count	Model Count	Model – of Zero Inflated	Count Model	Model of Zero Inflated
Intercept	Estimation Standard Error p – Value	0.6217 0.040 0.19	0.391 0.019 0.398	1.20 0.070 0.29	0.319 0.274 0.461	1.102 0.4 0.189
Gender of Insurers	Estimation Standard Error p – Value	-0.079 0.160 0.019	-0.0431 0.143 0.022	-0.1 0.20 0.009	-0.025 0.391 0.012	-0.120 0.49 0.0059
Age > 30	Estimation Standard Error	0.1701 0.31	0.438 0.30	0.001 0.200	0.103 0.40	-0.37 0.51

P-ISSN: 2204-1990; E-ISSN: 1323-6903
DOI: 10.47750/cibg.2023.29.03.017

Constraints		**Poisson**	**ZIP**		**ZINB**	
	p – Value	0.000	0.000	0.000	0.000	0.000
Age > 40	Estimation	-0.059	-0.032	-0.089	-0.02	-0.2
	Standard Error	0.012	0.31	0.198	0.171	0.20
	p – Value	0.389	0.0971	0.040	0.05	0.02
Age > 50	Estimation	-0.58	-0.798	-0.89	0.001	-0.109
	Standard Error	0.0251	0.0201	0.147	0.20	0.41
	p – Value	0.005	0.08	0.058	0.059	0.049
Age > 60	Estimation	0.599	0.39	0.075	0.37	0.090
	Standard Error	0.03	0.0198	0.10	0.15	0.35
	p –Value	0.09	0.068	0.064	0.070	0.053
Age > 70	Estimation	0.1601	0.462	0.200	0.26	0.089
	Standard Error	0.05	0.038	0.171	0.20	0.190
	p – Value	0.031	0.050	0.038	0.027	0.02
Age < 80	Estimation	0.057	0.0128	-0.08	0.017	-0.0449
	Standard Error	0.034	0.020	0.019	0.018	0.02
	p – Value	0.08	0.077	0.070	0.068	0.078
Exposure	Estimation	0.4217	0.6162	0.28	0.362	0.22
	Standard Error	0.030	0.021	0.024	0.017	0.02
	p – Value	0.048	0.060	0.051	0.058	0.044
Family Member	Estimation	-0.532	-0.801	-0.81	-0.4	-0.449
	Standard Error	0.051	0.029	0.198	0.167	0.186
	p – Value	0.041	0.03	0.029	0.040	0.030

Note: If the p–value <0.05, is significant, otherwise insignificant.

The table-4, signifies the regression analysis model of estimations with constraints Poisson ZIP and ZINB of model count and model of zero-inflated used estimations and standard value while checking the significance of p-values along with regression. For Poisson regression only model count, whereas ZIP regression model (Zero–inflated Poisson) count and of zero-inflated model were considered. Finally, for the Zero– inflated negative binomial regression) ZINB both count model and model of zero inflated taken in to analysis. The constraints used in the above analysis is gender, age of insurers, exposure and family members with a significant level of ($\alpha = 0.05$) P-value.

II) Health Insurance Claim Analysis of Female or male and Transgender:

The table-5 depicts descriptive statistics showing categorical variables on the parameters of claim frequency, family members and exposure among female or male, transgender and others, where the data transformed from 1 to 0 as displayed in table-1.

Table 5: Descriptive statistics showing Categorical Variables:

Parameters	Minimum	Maximum	Modus	Total
Frequency Claim	0	9	0	9201

P-ISSN: 2204-1990; E-ISSN: 1323-6903
DOI: 10.47750/cibg.2023.29.03.017

Members of Family	1	5	1	68765

Variable	Minimum	Maximum	Median	Total
Exposure	0	1	0.54317	19884.3

The table-6 portrays descriptive statistics showing categorical variables on the parameters of gender, various age groups from 40 – 80 years and above. where the data transformed from 1 to 0 as displayed in table-1.

Table 6: Categorical Variables Descriptive statistics:

Parameters	Categories	Explanation	Percentage
Gender	0 1	Female or Male Transgender	95.75% 3.14%
Age < 40	0 1	Others Age ≤ 30	59.64% 40.42%
Age < 50	0 1	Others Age 30 – 40	80.58% 19.40%
Age < 60	0 1	Others Age 40 - 50	81.40% 17.59%
Age < 70	0 1	Others Age 50 – 60	88.05% 11.89%
Age < 80	0 1	Others Age 60 - 70	90.50% 8.04%
Age >80	0 1	Others Age 70	96.08% 2.99%

From table-6, inferred that the numbers of transgender intricate in health insurance claim are very low compared to others (females and males). Furthermore, it also evidently specifies that the statistics of transgender of age more than 80 years age is smaller, infers transgenders of more than 80 years age shown less interest for insurance. The table- 7 illustrates the binomial regression estimations for (Po), (ZIP) and (ZINB)

Table 7: Estimation Model for Poisson, ZIP and ZINB regression:

Constraints		**Poisson**	**ZIP**		**ZINB**	
		Model Count	Model Count	Model – of Zero Inflated	Count Model	Model of Zero – Inflated
Intercept	Estimation Standard Error p – Value	0.579 0.049 0.0179	0.699 0.044 0.191	0.859 0.046 0.17	0.539 0.069 0.172	0.697 0.089 0.07

P-ISSN: 2204-1990; E-ISSN: 1323-6903
DOI: 10.47750/cibg.2023.29.03.017

Constraints		Poisson	ZIP		ZINB	
Gender of Insurers	Estimation	-0.011	-0.069	-0.079	-0.029	-0.069
	Standard Error	0.047	0.039	0.085	0.077	0.169
	p – Value	0.029	0.039	0.026	0.032	0.021
Age < 40	Estimation	0.299	0.249	-0.001	0.309	-0.019
	Standard Error	0.038	0.038	0.073	0.059	0.169
	p – Value	0.034	0.039	0.035	0.03	0.01
Age < 50	Estimation	-0.018	-0.029	-0.039	-0.023	-0.019
	Standard Error	0.049	0.041	0.078	0.069	0.173
	p – Value	0.019	0.028	0.019	0.019	0.016
Age < 60	Estimation	-0.751	-0.117	-0.257	0.048	-0.11
	Standard Error	0.018	0.016	0.061	0.05	0.078
	p – Value	0.049	0.049	0.049	0.047	0.049
Age < 70	Estimation	0.498	0.268	0.069	0.29	0.058
	Standard Error	0.010	0.008	0.032	0.027	0.034
	p – Value	0.055	0.059	0.055	0.05	0.049
Age < 80	Estimation	0.850	0.1501	0.036	0.12	0.06
	Standard Error	0.13	0.09	0.2	0.077	0.098
	p – Value	0.006	0.012	0.006	0.0060	0.006
Age > 80	Estimation	0.008	0.048	-0.079	0.014	-0.2
	Standard Error	0.082	0.065	0.069	0.065	0.087
	p – Value	0.10	0.18	0.107	0.12	0.08
Exposure	Estimation	0.2430	0.219	0.158	0.229	0.58
	Standard Error	0.086	0.079	0.17	0.08	0.12
	p – Value	0.08	0.09	0.069	0.076	0.060
Family Members	Estimation	-0.279	-0.19	-0.3	-0.398	-0.238
	Standard Error	0.08	0.089	0.198	0.147	0.175
	p – Value	0.03	0.040	0.020	0.029	0.02

Note: If the p – value < 0.05 is significant, otherwise insignificant.

Table – 7 explicates the estimation model for three types of regression parameters based on count and zero inflation influenced by age, gender, exposure, family members and claims with a significant level of α = 0.05 p value. The initial model fitting comparison is done with the help of Akaike Information Criterion (AIC) and -2log – likelihood, as shown in Table 8.

Table 8: Health Insurance Data using Model Fit Comparison

		Poisson	ZIP	ZINB
Female and Male	**AIC**	40987.29	28978.48	28988.35
	-2Log - likelihood	40224.26	26975.48	26112.17
Transgender and Other	**AIC**	39817.46	26112.44	24915.81
	-2Log - likelihood	39958.45	24332.44	24982.12

P-ISSN: 2204-1990; E-ISSN: 1323-6903
DOI: 10.47750/cibg.2023.29.03.017

Table-8 expound the model fit comparison of health insurance data using ACI value and -2Log likelihood models. ZIP model of regression fitted aptly for the data than other models. In Vuong non-nested test also applied for these models. The results of Vuong Non - nested tests using Health Insurance data showed in table-9 with female and male, transgender and other. The two models compared with two variables to test the significance includes Poisson and ZIP, ZIP and ZINB. Finally check the best preferred model at the end.

Table 9: Health Insurance data Results on Vuong Non - nested tests

	Comparison of Model	Statistic of Vuong Test	P value	Model Preference
Female and Male	**ZIP vs Poisson.**	-19.59	<0.02	ZIP
	ZINB vs ZIP	-1.18	<0.02	ZINB
Transgender and Other	**ZIP vs Poisson**	-18.52	<0.02	ZIP
	ZINB vs ZIP vs	-1.95	<0.02	ZINB

Discussions and Conclusion:

The 3 regression models are examined and utilised to fit the data from health insurance claims. As the standard error tends to be lower for the worst model, Tables 4 and 7 demonstrate that the Poisson regression model's standard error is lower than that of the other two models. The existence of extra zeros in the dataset explains why model does not appropriate the data well for the Poisson model, as can be seen from the conclusion. Then, ZIP and ZINB models are used to do the regression analysis for health insurance data claims in India. The tests for model fitting are run after fitting all three regression models. In Table 8 the results infer for model fit comparison test using AIC and -2loglikelihood for the models.

The AIC value is the highest in poisson model of regression indicating a subpar fit to the data. ZIP and ZINB, the final two models in each case, had decreased AIC values, representing a best suitable. However, because of the dataset's additional zeros and excessive dispersion of the other sections, the ZINB regression model fits somewhat better than the ZIP regression model, with a marginally lower value for AIC. To further compare the aforementioned models, the Vuong tests are completed. With of -20.47 value of test Vuong statistic with 0.02 p-value for the case one and for second case the value is -19.43 with 0.02 p-value, the Poisson and ZIP regression models have associated for the first time in each of the two situations. for the second case, representing that the ZIP model is more desirable.

Further advantageous model formerly contrasted to next mentioned model. Following assessment secondly of the ZIP and ZINB regression models, the ZINB model was initiate to be higher to the ZIP model in both instances, with a Vuong test statistic of -2.19 and a p-value of 0.01, respectively. The ZINB regression model is the best fitting and most recommended among the zero- inflated poisson regression models in the existence of additional zero and exceedingly distributed counts, according to the overall study and testing of fitted models.

P-ISSN: 2204-1990; E-ISSN: 1323-6903
DOI: 10.47750/cibg.2023.29.03.017

References:

1. Kibria, B. G. (2006). Applications of some discrete regression models for count data. *Pakistan Journal of Statistics and Operation Research*, 1-16.
2. Ozmen, I., & Famoye, F. (2007). Count regression models with an application to zoological data containing structural zeros. *Journal of Data Science*, *5*(4), 491-502.
3. Zulkifli, M., Ismail, N., & Razali, A. M. Zero-inflated Poisson versus zero-inflated negative binomial: Application to theft insurance data, The 7th IMT-GT International Conference on Mathematics, Statistics and its Applications (ICMSA 2011).
4. Ismail, N., & Zamani, H. (2013). Estimation of claim count data using negative binomial, generalized Poisson, zero-inflated negative binomial and zero-inflated generalized Poisson regression models. In *Casualty Actuarial Society E-Forum* (Vol. 41, No. 20, pp. 1-18).
5. Xu, L., Paterson, A. D., Turpin, W., & Xu, W. (2015). Assessment and selection of competing models for zero-inflated microbiome data. *PloS one*, *10*(7), e0129606.
6. Yang, S., Lisa L. Harlow, Gavino Puggioni and Colleen A. Redding (2017). A comparison of different methods of zero-inflated data analysis and its application in health surveys. Journal of Modern Applied Statistical Methods, Vol. 16, No. 1, pp. 518-543.
7. Astari, D. W. and Kismiantini (2019, October). Analysis of Factors Affecting the Health Insurance Ownership with Binary Logistic Regression Model. In *Journal of Physics: Conference Series* (Vol. 1320, No. 1, p. 012011). IOP Publishing.