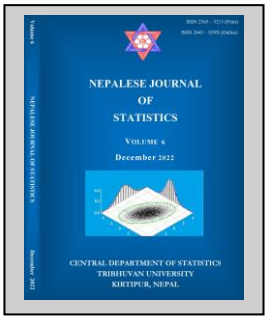
On the Use of Logistic Regression Model and its Comparison with Log-binomial Regression Model in the Analysis of Poverty Data of Nepal

Krishna Prasad Acharya[1], Shankar Prasad Khanal[2*] and Devendra Chhetry[3]



Submitted: 17 December 2022; Accepted: 25 December 2022
Published online: 27 December 2022
DOI: 10.3126/njs.v6i01.50806

ABSTRACT

Background: Previous literatures have indicated that log-binomial regression model is an alternative for the logistic regression model for frequent occurrence of event of outcome. The comparison of the performance of these two models has been found with reference to clinical/epidemiological data. Nonetheless, the application of log-binomial model and its comparison with the logistic model for poverty data has not been described.

Objective: To compare logistic and log-binomial regression model in terms of variable selection, effect size, precision of effect size, goodness of fit, diagnostics, stability of the model, and the issue of failure convergence.

Materials and Methods: Cross sectional data of 5988 households of Nepal Living Standard Survey 2010/11 has been used for the analysis. The performance of logistic and log-binomial model has been compared in terms of variable selection, effect size, and its precision for each covariate, goodness of fit using Hosmer - Lemeshow (H-L) test, diagnostics of the model, stability of the model using bootstrapping method, and the issue of failure convergence.

Results: Logistic model overestimates the effect size, yields wider 95% confidence interval than that of log - binomial model for each covariate. The greater elevation in risk for covariates varies from 13% to 173%. Logistic model satisfies goodness of fit of the model ($p = 0.534$), diagnostics tests, and stability of the model. However, log-binomial model grossly violates the goodness of fit of the model ($p = 0.0004$) but satisfies the model diagnostics and stability criteria.

Conclusion: Log-binomial model satisfies all criteria for model development and diagnostics except gross violation in goodness of fit of the model. However, logistic regression model satisfies all the criteria including goodness of fit of the model. On the basis of the entire comparison of model performance, logistic regression model is better fitted than the log-binomial model in fitting the poverty data set of Nepal.

Keywords: Diagnostics, elevation in risk, goodness of fit, log-binomial, logistic, poverty, stability, variable selection.

Address correspondence to the authors: Central Department of Statistics, Institute of Science and Technology, Tribhuvan University, Kirtipur, Kathmandu, Nepal.
Email: acharyakrishna20@gmail.com[1]; drshankarcds@gmail.com[2*](corresponding author email); chhetrydevendra@gmail.com [3]

INTRODUCTION

The logistic regression model is being used as a common method to study the associations of independent variables with the categorical dichotomous outcomes. Its use is frequent in case control, cohort studies and clinical trials. It has also been used in cross-sectional studies (Barros & Hirakata, 2003). It does not only measure the association of outcome variable with the independent variables, but also help to quantify the effect of these variables on the response variable. Logistic regression yields both regression coefficient for each independent covariate and the odds ratio (OR) based on the regression coefficient itself. Odds ratios are commonly reported in the analysis of different studies under such scenario (Davies, Crombie & Tavakoli, 1998) and seem to be relatively more appealing and effective in interpretations compared to regression coefficients. A considerable number of previous studies also indicated the use of relative risk, risk ratio (RR), prevalence ratio (PR), or rate ratio under such scenario (Davies et al., 1998; Holcomb, Chaiworapongsa, Luke & Burgdorf, 2001; Martinez, Leotti, Silva, Nunes, Machado & Corbellini, 2017; Gallis & Turner, 2019). There is still an academic debate regarding the issue of reporting which one either 'OR' or 'RR' is better. Some authors favor to report OR, some favor RR. Walter (1998), Olkin (1998), Newman (2001), and Cook (2002) favor to report odds ratio (OR) as they claimed that it is symmetric with the outcome. On the other hand, Sackett, Deeks and Altman (1996), De Andrade and Carabin (2011), and Gallis and Turner (2019) favor to use relative risk (RR) claiming that it is easily understandable. Lee (1994) has also remarked that odds ratio has been described as incomprehensible. Williamson, Eliasziw, & Fick (2013) has encouraged to use relative risk in epidemiological studies wherever possible, and to advocate its use. The odds ratios and the risk ratios are closer if the outcome of interest is very rare i.e. generally considered as less than 10 % (Greenland & Thomas, 1982; Greenland, Thomas & Morgenstern, 1986; Viera, 2008). If the outcome of interest is common (i.e. ≥10%), odds ratio will not be able to approximate risk ratio (Greenland, 1987; McNutt, Xiaonan Xue & Hafner, 2003; Katz, 2006; Viera, 2008; Ranganathan, Aggarwal & Pramesh, 2015; Gallis & Turner, 2019).

Initially, Wacholder (1986) recommended a simple approach of estimating risk ratios (RR) directly for studying the association of number of independent variables with the dichotomous response variable. Later, Barros and Vânia (2003) declared that generally the OR overestimates the RR in cross-sectional studies having frequent occurrences of event of interest. The basis of the log-binomial model is a generalized linear model with log link and binomial probability distribution, which results in risk ratio (RR). Robbins, Chao and Fonseca (2002), and McNutt et al. (2003) also highlighted their descriptions and applications. After that, Blizzard and Hosmer (2006) proposed the goodness of fit test and some diagnostics of the log- binomial regression model. There are

established methods to convert odds ratios into risk ratios. However, Robbins et al. (2002) had clearly indicated that these converted methods yielded inaccurate confidence intervals of estimates. It is also reported that there is failure convergence of log - binomial regression model for some applications (Williamson et al., 2013). The odds ratio and the relative risk can be computed in a different approach. The established technique for computing OR and RR in bivariate analysis is summarized in Table 1.

Table1. Layout of computation of OR and RR.

Independent variable	Outcome variable		Total
	Present	Absent	
Group I	a	b	n_{10}
Group II (Reference category)	c	d	n_{20}
Total	n_{01}	n_{02}	n

With reference to table1, probability of occurrence of 'a' is $p_1 = \frac{a}{n_{10}}$, and its complementary probability is $(1-p_1)$ in Group I. Similarly the probability of occurrence of 'c' in Group II is denoted by $p_2 = \frac{c}{n_{20}}$, and its complementary probability is $(1-p_2)$.

The odds ratio for the presence of outcome is defined as:

$$OR = \frac{p_1/(1-p_1)}{p_2/(1-p_2)}$$

The relative risk for the presence of defined outcome is simply defined as: $RR = p_1 / p_2$

The odds ratio is the ratio of two odds whereas the risk ratio is the ratio of two probabilities. The value of OR suppose is 3, is interpreted as the odds of having the outcome is 3 times higher in Group I than the reference group. If the probability of occurrence of outcome in Group I is 0.9 and 0.3 in reference group respectively, then risk ratio is interpreted as the group I is thrice as likely to have the outcome as the reference group. There is still some confusion while interpreting the odds ratios and relative risks which had been well indicated by Schwartz, Woloshin and Welch (1999); Zocchetti, Consonni and Bertazzi (1995). Further, Holcomb et al. (2001), and Baicus (2003) also clearly indicated the misinterpretation of odds ratio as risk ratio in considerable number of published articles in medical journals. However, the interpretation of RR and OR is not the focus of this paper; these issues have been discussed for the sake of completeness of the paper. Poverty is a complex issue and it possesses broadly two dimensions such as monetary poverty and non-monetary poverty. The analysis of this paper is exclusively focused on monetary poverty of Nepal. There are still 18.5% of households under poverty based on data of Nepal Living Standard Survey (2010/11) (Acharya, Khanal & Chhetry, 2022).

Identification of important factors associated with poverty using appropriate statistical model plays very important role for policy point of view. This is an attempt to recommend a more suitable model by comparing logistic and log-binomial regression model based on different criteria.

On extensive review of literature, it is found that the use of log- binomial regression is almost rare in social science related data problems especially on poverty data. The comparison of logistic regression model and the log-binomial regression model is also found quite rarely in social science researches. The objective of this paper is to apply the log-binomial regression model to the variables of the poverty data set of Nepal Living Standard Survey (2010/11), and to compare the results of logistic regression model and log-binomial regression model in terms of selection of variables in the final model, estimates, precision of the estimates, goodness of the fit of the model, diagnostics of the model, issue of convergence of the model, and stability of the model in the context of household poverty of Nepal. The issues of variable selection and model comparison are based on the paper by Acharya et al. (2022).

METHODOLOGY

Data

The study is based on cross sectional data of 5988 households of Nepal extracted from Nepal Living Standard Survey 2010/11. The survey was conducted by Central Bureau of Statistics (CBS) Government of Nepal. The un-weighted data was used for both log-binomial and logistic regression model. The detail survey methodology of Nepal Living Standard Survey is explained in survey report (CBS, 2011). The response variable for the model is the poverty status of household (coded 0 for non-poor and 1 for poor). Based on extensive review of relevant literature, altogether seven independent variables namely sex of household head (male vs. female), literacy status of household head (literate vs. illiterate), remittance receiving status(yes vs. no), land holding status(yes vs. no), access to market(better vs. poor), number of children in the household under 15 years of age (≤ 2 vs. > 2), and number of literate members of working age(≥ 1 vs. none) are considered initially for log-binomial regression model as used in logistic regression model. All details about selection of variables and need of dichotomization of independent variables, etc. had been described in Acharya et al. (2022).

Statistical model

The log-binomial regression model is a special type of generalized linear model for which the link function is log link. Logistic regression model is also a type of generalized linear model with logit link function. The response variable for both the models is dichotomous type.

The log-binomial regression model for p number of covariates $(X_1, X_2, \ldots\ldots\ldots\ldots, X_p)$ in association with binary response variable is given by:

$$\log \pi = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p \qquad (1)$$

where $\pi = \Pr ob[Y = 1 | X] = \exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p)$ for binary outcome Y, $\beta_1, \beta_2, \ldots\ldots\ldots\ldots, \beta_p$ are the regression coefficients for covariate

$(X_1, X_2, \ldots\ldots\ldots\ldots, X_p)$, β_0 is the constant term in the model. The link function for this model is log link. In this model RR can be computed as e^{β_j} for each considered covariate as done in the case of computing OR in logistic regression model. Regression coefficients are estimated by using the maximizing the likelihood function (for detail, please see McCullagh and Nelder, 1989)

$$l(\beta) = \sum_{i=1}^{n} \left[y_i \log \pi_i + (1 - y_i) \log (1 - \pi_i) \right] \qquad (2)$$

where, $\pi_i = e^{\left(\sum_{j=0}^{p} \beta_j x_{ij} \right)}$

Goodness of fit test and diagnostics of the fitted model

The goodness of fit of the log - binomial regression model can also be assessed by using Hosmer and Lemeshow (H-L) χ^2 test with (10 - 2 = 8) degrees of freedom using the formula

$$\hat{c} = \sum_{j=0}^{1} \sum_{k=1}^{10} \frac{\left(o_{jk} - \hat{e}_{jk} \right)^2}{e_{jk}} \qquad (3)$$

The observed and expected value in H-L χ^2 test in case of log-binomial regression model appear approximately equal but not exact. However, this test can also be applied for assessing the goodness of fit of the log-binomial regression model (Blizzard & Hosmer, 2006). The diagnostics of the fitted log-binomial regression model has been assessed graphically through the plot of (i) leverage in the vertical axis and the fitted model in the horizontal axis, and (ii) Chi-square displacement generated by the log- binomial model in the vertical axis and model fitted values in the horizontal axis (Blizzard & Hosmer, 2006).

The stability of the developed model has been evaluated by using bootstrapping method (Chen & George, 1985; Altman & Anderson, 1989; Saurbrei & Schumacher, 1992) as used for assessing the stability of Cox regression model. Same approach as applied by Khanal, Sreenivas & Acharya (2019) for comparing the stability of Cox proportional hazards model and two accelerated failure time models has been used to assess the stability of the logistic and log- binomial regression model. After fitting the final log-binomial regression model, the risk assessment of factors has been performed in terms of RR running the log-binomial model with same response variable on newly generated variable (X_i for i =1, 2, 3,…, p), where 1 represents the presence of any one factor,2 for presence of any 2 factors, and finally the presence of all factors in the final model respectively. The values of RR obtained from log-binomial model and the values of OR computed in similar manner from logistic regression model are compared.

Finally, the logistic regression model developed in the same data set by Acharya et al. (2022) and the log-binomial regression model developed in this attempt are compared in terms of variable selection, estimates, precision of estimates, goodness of fit of the model, regression diagnostics, stability of the model, and model convergence issue. Bootstrapping procedure has been done by

using R software, and remaining all statistical analysis has been performed by using STATA version 13.0.

RESULTS

There are altogether seven independent covariates associated with outcome variable used in the model. In order to identify the candidate variables for the final log-binomial regression model, simple log-binomial regression model considering one independent variable at a time separately is performed. Among these seven variables, only six variables; literacy status of household head - literate vs. illiterate (RR: 2.31, $p < 0.001$), remittance receiving status - yes vs. no (RR: 1.38, $p < 0.001$), land holding status - yes vs. no (RR: 1.79, $p < 0.001$), access to market - better vs. poor (RR: 2.25, $p < 0.001$), number of children under 15 years of age-less than or equal to 2 vs. more than 2 (RR: 3.68, $p < 0.001$, and number of literate members of working age- at least one vs. none (RR: 1.97, $p < 0.001$), each has come out significantly associated with response variable at 5% level of significance except sex of household head - male vs. female (RR: 0.92, $p = 0.195$). Though sex of household head has not come out statistically significant even in the simple log-binomial regression model, it is also considered as one of the candidate variables for developing multiple log-binomial regression model treating it as a known confounder. Hence, the seven variables including sex of household head are considered as potential candidate variables for the final multiple log-binomial regression model.

Results of multiple log-binomial regression model

In order to select the significant variables in the final model both stepwise forward selection and backward selection procedure are adopted considering seven candidate variables. Both selection procedures have yielded six common set of significant variables at 5% level of significance except sex of household head. The values of RR, standard error (S.E.) of RR, p-values and 95% CIE for each independent factor associated with poverty obtained through multiple log- binomial regression model are presented in Table 2.

Among finally selected six independent predictors associated with poverty, the risk of household having more than two children under 15 is found to be the highest (RR: 2.96, 95% CIE: 2.66, 3.28) followed by household having illiterate household head, household with poor access to the market center, household not receiving remittance, household with no land respectively. The risk of household being poor among household not having single literate members of the working age in comparison with households having at least one literate member is found to be the least (RR: 1.16, 95% CIE: 1.05, 1.29). This can be interpreted as the household not having single literate members of the working age is 1.16 times as likely to have poorer than those household having at least one literate members of the working age. The goodness of fit of the model assessed by H-L (χ^2) test with 8 degrees of freedom is highly violated ($p = 0.0004$) (Table 2).

Table2. Results of multiple log - binomial regression model.

Independent variables	RR	S.E.	p-value	95% CIE	
Literacy status of household head:					
Literate	1.00				
Illiterate	1.68	0.1006	< 0.001	(1.49	1.89)
Remittance receiving status:					
Yes	1.00				
No	1.45	0.0685	< 0.001	(1.33	1.59)
Land holding status:					
Yes	1.00				
No	1.22	0.0594	< 0.001	(1.11	1.34)
Access to market:					
Better	1.00				
Poor	1.51	0.0888	< 0.001	(1.34	1.69)
Number of children under 15:					
≤ 2	1.00				
> 2	2.96	0.1590	< 0.001	(2.66	3.28)
No. of literate members of working-age:					
≥ 1	1.00				
0	1.16	0.0606	< 0.001	(1.05	1.29)
Constant	0.05	0.0034	< 0.001	(0.05	0.06)
Log likelihood (only with intercept) = - 4068.888; Log likelihood (full model) = - 2412.336					
AIC = 0.808; BIC = - 47195.150; H-L (χ^2) with 8 d. f. = 28.602, p = 0.0004					

Source: computed based on NLSS 2010/11 data

Results of diagnostics for log-binomial regression model

The plot of the leverage values in y-axis and the model fitted values in x-axis is presented in Figure1 (a). One data value seems to be in the top right corner having relatively greater leverage than others. Majority of the leverage values are found less than 0.008 and the extreme one leverage in this dataset is also less than 0.01. Hence, all the leverage values are found to be less than 0.08 which indicates that there is not violation of the diagnostics of the model assessed based on leverage (Blizzard & Hosmer, 2006).

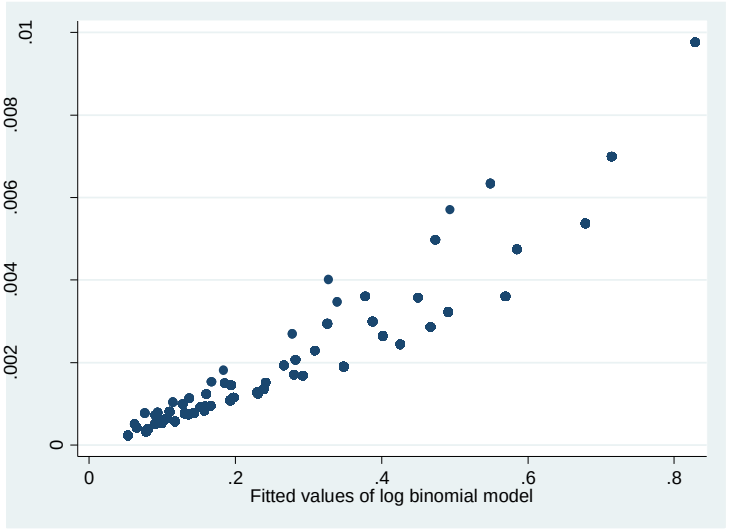


Fig. 1(a). Leverage and fitted values of log - binomial model.

The diagnostic of the fitted model has also been assessed through the plot keeping $\Delta\chi^2$ in vertical axis and the values of fitted log-binomial regression model in horizontal axis with plotting symbol proportional to Cook's distance (Figure1(b)). There are four poorly fit data points with $\Delta\chi^2$ > 10 (Blizzard & Hosmer, 2006). The circles of these four data points are observed to be larger than others. Two data points are lying in a bit far away and another one is farther away from others in the right lower corner. There is not much serious violation of the diagnostics of the fitted model evaluating on the basis of the plot of $\Delta\chi^2$ vs. fitted log - binomial model.

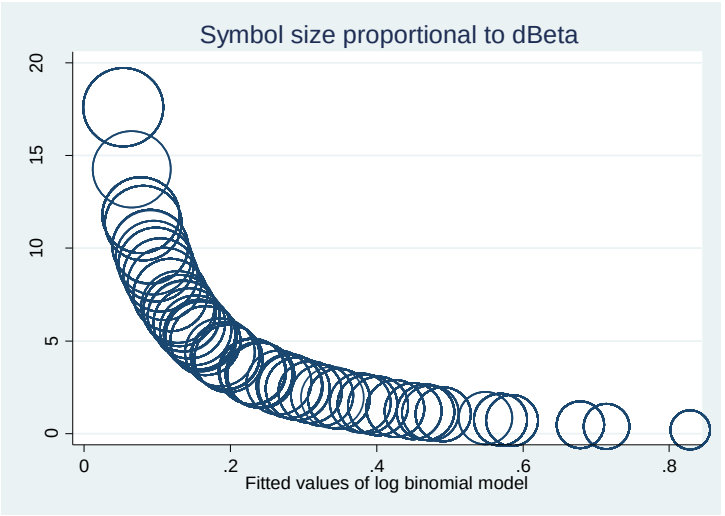


Fig. 1(b). Graph of $\Delta\chi^2$ vs. values of fitted log - binomial model with plotting symbol proportional to Cook's distance.

Comparison of logistic and log-binomial regression model

Model building of both logistic and log-binomial regression models are started taking with same set of seven covariates. Out of these seven covariates both models have come up with six significant covariates except variable 'sex of household head'. While comparing the effect size(OR for logistic regression and RR for log - binomial regression model) for each covariate, logistic regression model overestimates the effect size (Table3) and wider width of 95% confidence interval estimation than that of log-binomial regression model (Table 3). Wider confidence interval estimation of effect size for each covariate in logistic regression model clearly indicates the lesser precision of the estimate than that of log-binomial regression model. The value of OR for each independent variable obtained from logistic regression model overestimates the value of RR obtained from log-binomial regression model. The reference value for each OR and RR is 1.

The elevation of risk percentage for a covariate is computed as $\text{Elevation risk}(\%) = \left[(\text{OR-1}) - (\text{RR-1})\right] \times 100$. For example; let us consider the elevation risk (%) for the variable literacy status of the household head is computed with reference to Table3 as $\left[(2.20 - 1) - (1.68 - 1)\right] \times 100 = (1.20 - 0.68) \times 100 = 52.0\%$. It is computed in similar fashion for other covariates (Table3). The elevation of risk for the covariate estimates generated by logistic regression varies from 13% to 173% than those generated by log-binomial regression model. The highest elevation of risk (173%) is noted for the variable 'number of children less than 15 years of age' and the least elevation of risk (13%) is observed for the variable 'number of literate members of working age'.

Logistic regression model has satisfied the goodness of fit of the test as assessed by H-L (χ^2) test (8 d.f.) with non-significant result (p = 0.534) whereas the goodness of fit test of the log-binomial regression model as assessed by H-L (χ^2) test (8 d.f.) is grossly violated (p = 0.0004). The value of AIC is lees, and the value of BIC is greater with negative sign in log-binomial model compared to logistic regression model. Neither logistic regression nor log-binomial regression model has faced the model failure convergence i.e. both the models do not show the misbehavior.

Table 3. Comparison of logistic and log - binomial regression model in terms of variable selection, estimates, precision of the estimates, goodness of fit of the model, AIC and BIC.

Independent variables	Logistic regression model		Log - binomial regression model		Elevation in risk (%)
	OR(95% CIE)	Width of interval	RR(95% CIE)	Width of interval	
Literacy status of household head:					
Literate	1.00		1.00		
Illiterate	2.20 (1.86 2.61)	0.75	1.68 (1.49 1.89)	0.4	52
Remittance receiving status:					
Yes	1.00		1.00		
No	1.90 (1.64 2.20)	0.56	1.45 (1.33 1.59)	0.26	45
Land holding status:					
Yes	1.00		1.00		
No	1.53 (1.31 1.78)	0.47	1.22 (1.11 1.34)	0.23	31
Access to market:					
Better	1.00		1.00		
Poor	1.77 (1.52 2.07)	0.55	1.51(1.34 1.69)	0.35	26
Number of children under 15:					
≤ 2	1.00		1.00		
> 2	4.69(4.06 5.42)	1.36	2.96(2.66 3.28)	0.62	173
No. of literate members of working-age:					
≥ 1	1.00		1.00		
0	1.29(1.07 1.56)	0.49	1.16(1.05 1.29)	0.24	13
H-L (χ^2) with 8 d.f	6.05, p = 0.534		28.602, p = 0.0004		
AIC	4813.844		0.808		
BIC	4860.727		- 47195.150		

Source: Results of logistic regression are adopted from Acharya et al. (2022); Results of log-binomial regression are computed based on NLSS 2010/11 data.

Comparison based on diagnostics of the model

The diagnostics of the fitted logistic regression model was assessed graphically through the (i) plots of $\Delta\beta$ vs. model estimated probability, and (ii) $\Delta\chi^2$ vs. model estimated probability with symbol size proportional to $\Delta\beta$ (Figure2(a) and 2(b)). Both the figures have reasonably satisfied the diagnostics of the model through visual assessment except 2 data points greater than 1 in Figure2 (a), and the value of $\Delta\chi^2$ and $\Delta\beta$ are not influenced by covariate patterns except for one covariate (Figure2 (b)).

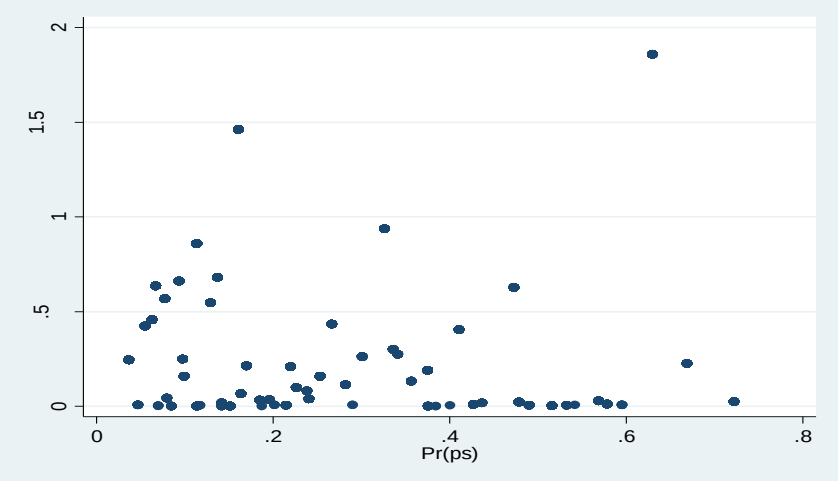


Fig. 2(a). Plot of $\Delta\beta$ versus logistic regression model estimated probability. (Source: Acharya et al. (2022)

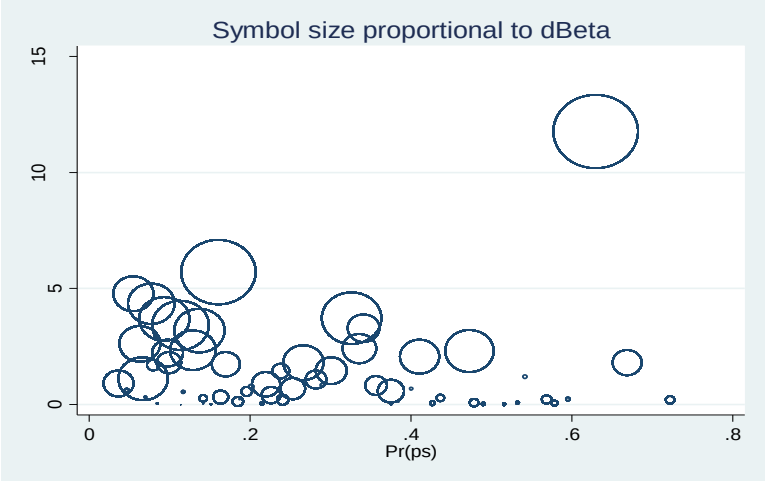


Fig. 2(b). Plot of $\Delta\chi^2$ versus logistic model estimated probability with symbol Size proportional to $(\Delta\beta)$. (Source: Acharya et al. (2022))

The diagnostics of the fitted log - binomial regression model is assessed graphically by (i) leverage versus predicted value of log - binomial regression model (Figure1 (a)) , and by (ii) graph of $\Delta\chi^2$ versus values of fitted log - binomial model with plotting symbol proportional to Cook's distance(Figure1(b)). Based on the visual assessment of the plots, the fitted log-binomial model (Figure 1(a) & (b)) reasonably satisfies the diagnostics of the model.

Comparison based on stability of the model

High repetition of each variable in each final model is assessed through bootstrapping resampling technique running each model 1000 times with final set of independent variables. The major objective of this method was to identify the importance of each variable in each final model through the maximum number of occurrences of each variable in each model. Naturally, the higher the repetition of occurrence of variable indicates the more importance in the model and consequently indicates the stability of the fitted model. In each model same five variables are found repeated 100% of times, and only one variable ' number of literate members of working age group' is repeated 97.4% of times. Hence, both models have satisfied the stability criteria indicating that the selected variables are almost equally important in each model.

Comparison based on risk assessment

The risk assessment has been performed for each model based on the presence of any one, any two risk factors, etc. by running logistic and log - binomial model separately. The risk of households being poor is found increasing continuously as the number of factors increases in each model (Figure3). However, logistic regression model overestimates the risk for each factor than that of log - binomial regression model analogous to the results of the original logistic and log - binomial model we used in the analysis.

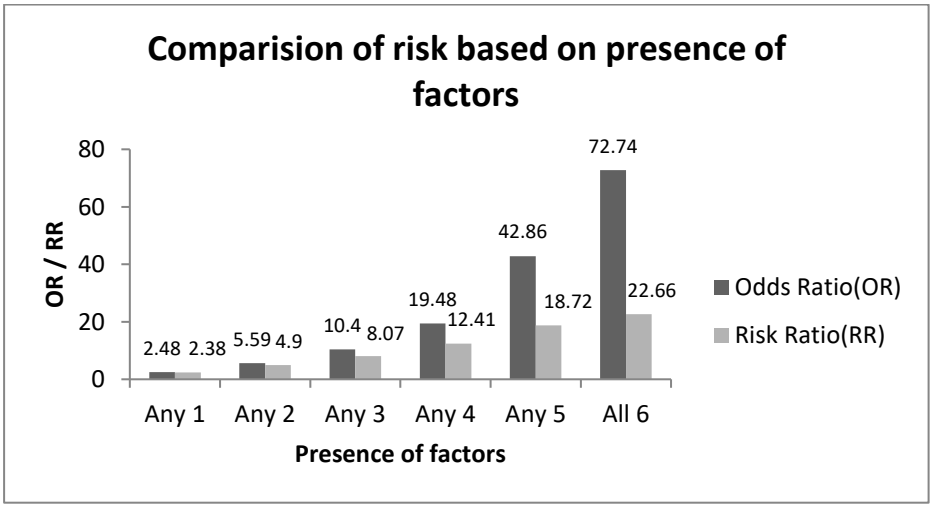


Fig. 3. Risk assseemsnt based on presence of factors for logistic and log-binomial regression model.

DISCUSSION

The findings of the study have clearly indicated that each model has picked up the same set of six independent predictors in the final model from the same pool of variables. Both stepwise forward and backward selection procedures have been applied to select the variables in each of the final model to akin whether different selection methods serve differently in each model. However, both selection procedures have behaved in a similar manner in each model by selecting the same

set of variables. The effect size of each independent variable is overestimated in logistic regression model as compared to that of log-binomial regression model. Similar findings have been reported by other studies (Barros and Hirakata, 2003; Espelt et al., 2017; Diaz-Quijano, 2012). In logistic regression model OR varies from 1.29 to 4.69, and in log- binomial regression model RR varies from 1.16 to 2.96. There is clear greater elevation of risk in logistic regression model as compared to log-binomial regression model for each independent variable, and it varies from 13% to 173%. While comparing effect size for each variable within the model and between the models, if it is smaller or larger in a variable in logistic model, it is also smaller or larger in log - binomial model for the same variable. Just for example; the highest OR value of 4.69 for a variable 'number of children under 15 years' in logistic regression model, and the highest RR value of 2.96 for the same variable in log - binomial regression model. The precision of effect size of each variable in log-binomial model is better than that of logistic regression model as assessed by 95% CIE.

There is remarkable wider interval width of effect size in each variable in logistic regression model than that of log-binomial regression model. This finding is similar to the findings of Deddens and Petersen (2008); Barr, et al. (2016). While comparing the goodness of fit of two models, logistic regression model has satisfied the goodness of fit criteria but log - binomial regression model has grossly violated as assessed by H-L (χ^2) test. The violation in this regard in log-binomial model might be because of considering only categorical independent variables, but the exact reason is not known. There is not any problem of failure convergence in both models. Some studies have reported the issue of failure convergence specially while running log-binomial regression model (Williamson, et al., 2013; Barros & Hirakata, 2003; De Andrade & Carabin, 2011). The value of AIC of log - binomial model is smaller than that of logistic model; the value of BIC is larger in magnitude in log - binomial model than that of logistic model but with negative sign. The diagnostics of the logistic model assessed through (i) the graph of $\Delta\beta$ versus model estimated probability, and (ii) the graph of $\Delta\chi^2$ versus model estimated probability with symbol size proportional to $(\Delta\beta)$ are reasonably satisfied. The diagnostics of the fitted log - binomial model assessed through graph of (i) leverage versus model fitted value, and (ii) the graph of $\Delta\chi^2$ versus model fitted value with symbol proportional to Cook's distance is also reasonably satisfied. Similar findings are reported by (Blizzard & Hosmer, 2006) regarding the regression diagnostics based on the comparison of these two models using follow up study of infants. However, our finding regarding the goodness of fit of the model for log - binomial model does not support the study of Blizzard & Hosmer (2006) but supports for the logistic regression model. While comparing the stability of the fitted model evaluated using bootstrapping resampling method of running each model 1000 times, all five variables are repeated 100% times except one variable's repetition of 97% of times. This signifies that both finally fitted models can be considered as stable.

Limitation

All independent variables used in both logistic and log - binomial model are of categorical type. The reasons behind the consideration of categorized variables are ease of interpretations of effect size and effective implementation in policy implications for comparing the groups such as advantaged vs. non advantaged groups, etc.

CONCLUSION

Both logistic and log- binomial model possess same behavior in terms of selection of variables in the final model, diagnostics of the fitted model, stability of the model and issue of failure convergence. However, logistic regression model overestimates the effect size, wider CIE of effect size than that of log-binomial model. The value of AIC is smaller in log- binomial model than that of logistic model. Comparison based on the estimates, precision of estimates, and AIC, log-binomial model is better than logistic regression model in this cross sectional poverty data of Nepal. Logistic regression model satisfies the goodness of fit but log - binomial model grossly violates. Logistic regression model is better than log - binomial regression model for this poverty data comparatively based on the entire comparison including goodness of fit of the model. Nonetheless, log - binomial model is a good alternative for logistic regression model, especially for not overestimating effect size and its better precision.

CONFLICT OF INTEREST

The authors declared absence of conflict of interest.

ACKNOWLEDGEMENTS

We would like to acknowledge Research Committee, Central Department of Statistics, TU for comments and suggestions, and to University Grants Commission Nepal for Ph.D. fellowship of this work as it is a part of Ph.D. research work. We also like to thank Prof. Leigh Blizzard, Menzies Institute for Medical Research, University of Tasmania for providing us STATA codes for computing H-L Chi-Square in log -binomial model, and would like to acknowledge unknown reviewers whose comments and suggestions have greatly helped to improve the manuscript.

REFERENCES

Acharya, K. P., Khanal, S. P., & Chhetry, D. (2022). Factors Affecting Poverty in Nepal - A Binary Logistic Regression Model Study. *Pertanika Journal of Social Science and Humanities*, *30*(2). doi: https://doi.org/10.47836/pjssh.30.2.12

Altman, D. G., & Anderson, P. K. (1989). Bootstrap investigation of the stability of a Cox regression model. *Statistics in Medicine, 8*(7), 771-783. doi: 10.1002/sim.4780080702

Baicus, C. (2003). Relative risks or odds ratios? *Canadian Medical Association Journal*, *168*(12), 1529.

Barr, Margo L., Clark, Robert, & Steel, D. G. (2016). *Examining associations in cross-sectional studies.* National Institute for Applied Statistics Research Australia, University of Wollongong. Retrieved from https://ro.uow.edu.au/niasrawp/35

Barros, A. J. D., & Hirakata, V. N. (2003). Alternatives for logistic regression in cross-sectional studies: An empirical comparison of models that directly estimate the prevalence ratio. *BioMed Central Medical Research Methodology, 3*(21). doi: https://doi.org/10.1186/1471-2288-3-21

Blizzard, L., & Hosmer, D. W. (2006). Parameter estimation and goodness-of-fit in log binomial regression. *Biometrical Journal*, *48*, 5–22. doi: 10.1002/bimj.200410165

Central Bureau of Statistics. (2011). *Nepal Living Standard Survey (2010/11). Statistical Report, Volume One.* Central Bureau of Statistics, National Planning Commission Secretariat, Government of Nepal.

Chen, C. H., & George, S. L. (1985). The bootstrap and identification of prognostic factors via Cox's proportional hazards regression model. *Statistics in Medicine, 4*(1), 39-46. doi: 10.1002/sim.4780040107

Cook, T. D. (2002). Advanced statistics: Up with odds ratios! A case for odds ratios when outcomes are common. *Academic Emergency Medicine*, *9*, 1430–1434. doi: 10.1111/j.1553-2712.2002.tb01616.x

Davies, H. T. O., Crombie, I. K., & Tavakoli, M. (1998). When can odds ratios mislead?. *British Medical Journal, 316*(7136), 989-991. doi: 10.1136/bmj.316.7136.989

De Andrade, B. B., & Carabin, H. (2011). On the estimation of relative risks via log binomial regression. *Revista Brasileira de Biometria, 29*(1), 15.

Deddens, J. A., & Petersen, M. R. (2008). Approaches for estimating prevalence ratios. *Occupational and environmental medicine, 65*(7), 501-506. doi: https://doi.org/10.1136/oem.2007.034777

Diaz-Quijano, F. A. (2012). A simple method for estimating relative risk using logistic regression. *BMC Medical Research Methodology, 12*(1), 1-6. doi: 10.1186/1471-2288-12-14.

Espelt, A., Marí-Dell'Olmo, M., Penelo, E., & Bosque-Prous, M. (2017). Applied prevalence ratio estimation with different Regression models: An example from a cross-national study on substance use research. *Adicciones, 29(*2), 105-112. doi: 10.20882/adicciones.823

Gallis, J. A., & Turner, E .L. (2019). Relative measures of association for binary outcomes: Challenges and recommendations for the global health researcher. *Annals of Global Health, 85*(1): 137, 1–12. doi: https://doi.org/10.5334/aogh.2581

Greenland, S., & Thomas, D. C. (1982). On the need for the rare disease assumption in case-control studies. *American Journal of Epidemiology, 116*(3), 547–553.

Greenland, S. (1987). Interpretation and choice of effect measures in epidemiologic analyses. *American Journal of Epidemiology, 125*(5), 761–768.

Greenland, S., Thomas, D. C., & Morgenstern, H. (1986). The rare-disease assumption revisited: A critique of "estimators of relative risk for case-control studies. *American Journal of Epidemiology*, *124*(6), 869–883.

Holcomb, W. L., Chaiworapongsa, T., Luke, D. A., & Burgdorf, K. D. (2001). An odd measure of risk: use and misuse of the odds ratio. *Obstetrics & Gynecology*, *98*(4), 685–688. doi: https://doi.org/10.1016/s0029-7844(01)01488-0

Katz, K. A. (2006). The (relative) risks of using odds ratios. *Archives of Dermatology*, *142*(6), 761–764.

Khanal, S.P., Sreenivas, V., & Acharya, S. K. (2019). Comparison of Cox proportional hazards model and lognormal accelerated failure time model: Application in time to event analysis of acute liver failure patients in India. *Nepalese Journal of Statistics, 3*, 21–40. doi: https://doi.org/10.3126/njs.v3i0.25576

Lee, J. (1994). Odds ratio or relative risk for cross-sectional data. *Int J Epidemiol*, *23*(1), 201–203. doi: https://doi.org/10.1093/ije/23.1.201

Martinez, B. A. F., Leotti, V. B., Silva, G. D. S. E., Nunes, L. N., Machado, G., & Corbellini, L. G. (2017). Odds ratio or prevalence ratio? An overview of reported statistical methods and appropriateness of interpretations in cross-sectional studies with dichotomous outcomes in veterinary medicine. *Frontiers in Veterinary Science, 4*, 193. doi: https://doi: 10.3389/fvets.2017.00193

McCullagh, P., & Nelder, J. A. (1989). *Generalized Linear Models.* Chapman & Hall

McNutt, L.-A., Xiaonan Xue C. W., and Hafner J. P. (2003). Estimating the relative risk in cohort studies and clinical trials of common outcomes. *American Journal of Epidemiology, 157*, 940–943. doi: 10.1093/aje/kwg074

Newman, S. C. (2001). Biostatistical Methods in Epidemiology (pp 35-40). New York: Wiley

Olkin, I. (1998). Letter to the editor. *Evidence-Based Medicine 3*, 71.

Ranganathan, P., Aggarwal, R., & Pramesh, C. S., (2015). Common pitfalls in statistical analysis: Odds versus risk. *Perspectives in Clinical Research, 6*(4), 222-224. doi: https://doi.org/10.4103%2F2229-3485.167092

Robbins, A. S., Chao, S. Y., & Fonseca, V. P. (2002). What's the relative risk? A method to directly estimate risk ratios in cohort studies of common outcomes. *Annals of Epidemiology, 12*, 452–454. doi: 10.1016/s1047-2797(01)00278-2

Sackett, D. L., Deeks J. J, & Altman, D. G. (1996). Down with odds ratios!. *Evidence Based Medicine*, *1*, 164–166. Retrieved from https://ebm.bmj.com/content/ebmed/1/6/164.full.pdf

Saurbrei, W., & Schumacher, M. (1992). A bootstrap resampling procedure for model building application to the Cox regression model. *Statistics in Medicine, 11*, 2093-2109. doi: https://doi.org/10.1002/sim.4780111607

Schwartz, L. M., Woloshin, S., & Welch, H. G. (1999). Misunderstandings about the effects of race and sex on physicians' referrals for cardiac catheterization. *New England Journal of Medicine, 341*(4), 279–283.

Viera, A. J. (2008). Odds ratios and risk ratios: What's the difference and why does It matter?. *Southern Medical Journal, 101*(7), 730-734. doi: https://doi.org/10.1097/smj.0b013e31817a7ee4

Wacholder, S. (1986). Binomial regression in GLIM: Estimating risk ratios and risk differences. *American Journal of Epidemiology, 123*, 174–184. doi: 10.1093/oxfordjournals.aje.a114212

Walter, S. (1998). Letter to the editor. *Evidence-Based Medicine*, *3*(71).

Williamson, T., Eliasziw, M., & Fick, G. H. (2013). Log-binomial models: exploring failed convergence. *Emerging Themes in Epidemiology, 10*(14). doi: https://doi.org/10.1186/1742-7622-10-14

Zocchetti C., Consonni D., and Bertazzi P. A. (1995). Estimation of prevalence rate ratios from cross-sectional data. *International Journal of Epidemiology, 24*(5), 1064–1067.

Reference to this paper should be made as follows:
Acharya, K. P., Khanal, S. P., & Chhetry, D. (2022). On the use of logistic regression model and its comparison with log-binomial regression model in the analysis of poverty data of Nepal. *Nep. J. Stat, 6,* 63-79.

