ProbStat Forum, Volume 08, October 2015, Pages 140–147
ISSN 0974-3235

ProbStat Forum is an e-journal.
For details please visit www.probstat.org.in

Comparison between count regression and binary logistic regression models in the analysis of adverse drug reaction data

Vasudeva Guddattu[a], Aruna Rao K[b], Thiyagu Rajkannan[c]

[a] *Department of Statistics, Manipal University, Manipal, India.*
[b] *Department of Statistics, Mangalore University, Mangalore, India.*
[c] *Department of Pharmaceutical Health Services Research, University of Maryland School of Pharmacy, Baltimore, USA.*

Abstract. Many researches prefer to categorize count variable as binary and do the analysis.By categorizing the count variables into binary variable,we are loosing the information. In this paper, we analyze the use of count regression models as compared to logistic regression analysis. We had considered a count data of adverse drug reactions and applied different types count regression models.Later logistic regression model was fitted by categorizing the count data into binary.From the analysis,zero inflated negative binomial regression model fits better than other count regression models and logistic regression.Count regression models will have a finer interpretation of regression coefficients as compared to logistic regression model.Zero inflated negative binomial regression model has an added advantage of modeling over dispersion and excess zeros.

1. Introduction

It is common practice among medical practitioners and researchers to convert count variable as a binary variable to identify the risk factors. Logistic regression is a common tool used to identify risk factors associated with binary variable. It is simple to understand and interpretation is straightforward in terms of odds ratios. In the process, some information is lost. In the present paper, we made an attempt to compare the results of logistic regression and count regression. Thiyagu (2010) has analysed the data on number of Adverse drug reaction(ADR) by converting it as a binary variable. Since the number of ADR is a count variable, we have analysed the same data using count regression. The independent variables considered in the model are age, gender, type of system involved and type of drug given. Initially we fitted Poisson, negative binomial (NB), generalized Poisson (GP), Zero inflated Poisson (ZIP), Zero inflated negative binomial (ZIP) and zero inflated generalized Poisson (ZIGP) for the data without covariates. Bayesian Information criteria (Schwarz, 1978) was used to select the best model among the listed models. The Bayesian Information criteria (BIC) was smaller for the ZINB regression model. In the second stage, we included covariates for selected model in the first stage. Backward elimination criteria was used for selection of variables using BIC. Finally, the ZINB regression model with age, gender, type of system involved and type of drug given as covariates emerge as the best. For the sake of completeness, a logistic regression model was also fitted using the backward elimination procedure. Further BIC was minimum for the ZINB regression model as compared to the logistic regression model. The model selected in final step using ZINB has one more variable than logistic regression model. ZINB regression model selects four variables as compared to three in the logistic regression model. In the ZINB model, interpretation of regression parameter can be made in terms

Received: 20 March 2014; Revised 25 July 2015, Accepted: 05 August 2015

Email addresses: vasudev.guddattu@gmail.com (Vasudeva Guddattu), arunaraomu@yahoo.com (Aruna Rao K), thiyagu.rajakannan@gmail.com (Thiyagu Rajkannan)

of relative risk or rate ratio where as in logistic regression, it is made in terms of odds ratios. This analysis demonstrated that count data with many outcomes needs to be analysed as a count regression model as it offers a finer interpretation of regression coefficients than a logistic regression model. Our results are in agreement with Greenland (2004), Mcnutt et al. (2003) where they show the superiority of the count regression model in estimating relative risk than the logistic regression.

The organisation of the paper is as follows. Section 2 describe various regression models. The description of dataset and the selection of the model for the count data is presented in section 3. Section 4 deals with the selection of regression models for logistic regression by converting count data to binary and section 5 discusses the choice between the models. Section 6 presents the concluding remarks.

2. Models for count data

Here we give a brief introduction to the regression models used to model count data. For the case study presented in section3, we used Poisson, negative binomial, generalized Poisson, Zero inflated Poisson, Zero inflated negative binomial and zero inflated generalized Poisson regression models.

2.1. Poisson and zero inflated Poisson regression

let $Y_i; \quad i = 1, 2..., n$ denote the count random variable.The Poisson regression model with single covariate x_i is defined as

$$Log(\lambda_i) = \beta_0 + \beta_1 x_i; \quad i = 1, 2..., n. \tag{1}$$

where λ_i denotes the mean of Poisson random variable Y_i. When we have excess zeros, zero inflated Poisson regression model is an alternative to poisson regression. The probability mass function of zero inflated Poisson regression is given by

$$P(Y_i = y_i) = \begin{cases} P + (1-P)e^{(-\lambda_i)} & y_i = 0 \\ \frac{(1-P)e^{(-\lambda_i)}(\lambda_i)^{y_i}}{y_i!} & y_i = 1, 2..., \lambda_i > 0, 0 < P < 1 \end{cases} \tag{2}$$

Here λ_i is regressed to an independent variable x_i as described in (1).

2.2. Negative binomial and zero inflated negative binomial regression

Let Y_i, $i = 1, ..., n$ be independent random variables with mean μ_i and probability mass function

$$g(Y_i = y_i, \mu_i, \phi, c) = \frac{\Gamma_{y_i + \phi^{-1}\mu_i^{1-c}}}{\Gamma_{y_i+1}\Gamma_{\phi^{-1}\mu_i^{1-c}}} \frac{\phi^{y_i}\mu_i^{cy_i}}{(1+\phi\mu_i^c)^{y_i+\phi^{-1}\mu_i^{1-c}}}; \quad \phi \geq 0, \mu_i > 0. \tag{3}$$

In the above mass function, μ_i, ϕ, c are respectively called mean, dispersion and index parameter. The probability distribution with pmf defined in (3) is the general form of negative binomial distribution. The commonly used distribution corresponds to the case of c=0, negative binomial 1 distribution (NB1); c=1 , negative binomial 2 distribution (NB2).
In NB1 distribution, mean and variance is linearly related by the relation

$$E(Y_i) = \mu_i, \quad V(Y_i) = \mu_i(1+\phi), \quad i = 1, ..., n \tag{4}$$

while for NB2 distribution, the relation is quadratic with

$$E(Y_i) = \mu_i, \quad V(Y_i) = \mu_i(1+\phi\mu_i), \quad i = 1, 2..., n \tag{5}$$

For the current study, we had used NB2 distribution.We restrict our attention only to this distribution and in the sequel we refer to it as negative binomial distribution.

The probability mass function of the zero inflated negative binomial distribution is given by

$$\begin{aligned} f(Y_i = y_i, P, \mu_i, \phi) &= P + (1-P)(1+\phi\mu_i)^{\frac{-1}{\phi}} & y_i = 0 \\ &= (1-P)\frac{\Gamma_{y_i+\phi^{-1}}}{\Gamma_{y_i+1}\Gamma_{\phi^{-1}}}\frac{\phi^{y_i}\mu_i^{y_i}}{(1+\phi\mu_i)^{y_i+\phi^{-1}}}; & y_i = 1, 2... \end{aligned} \tag{6}$$

The mean and variance for the pmf defined in (6) is given by

$$E(Y_i) = (1-P)\mu \quad V(Y_i) = (1-P)\mu_i(1 + P\mu_i + \phi\mu_i), \quad i = 1, 2, \ldots, n. \tag{7}$$

It is customary to refer μ_i, $i = 1, 2, \ldots, n$ as mean parameter and ϕ as dispersion parameter in zero inflated negative binomial distribution although, they do not correspond to mean and variance of these distributions. To develop a regression model based on zero inflated negative binomial distribution and negative binomial distribution, covariates $x_0, x_1, ...x_n$ are related to μ_i through log link function given by

$$Log(\mu_i) = \beta_0 x_{i0} + ... + \beta_k x_{ik}; \quad i = 1, 2, ..., n \tag{8}$$

where $\beta_0, \beta_1...\beta_k$ are the regression parameters of the model. To accommodate intercept term, x_{i0} is taken as 1, $i = 1, 2, ..., n$. Further for zero inflated negative binomial regression,we do not link covariates for inflate parameter P and dispersion parameter ϕ. The parameters of the model are estimated using maximum likelihood principle (Lehman and Romano, 2005).

2.3. Generalized Poisson and zero inflated generalized Poisson regression

The probability mass function of generalized Poisson distribution is given by

$$f(Y = y; \mu, \phi) = \frac{\mu(\mu + \phi\mu y)^{y-1} e^{-\mu-\phi\mu y}}{y!}, \quad y = 0, 1, 2, ... \tag{9}$$

where $\mu > 0$ and $max(-1, -\frac{\mu}{m}) < \phi\mu < 1$ with m a largest positive integer such that $1 + m\phi > 0$. This form of the generalized Poisson distribution is used by Consul (1989) and Gupta et al. (2004). For more details of this distribution see Johnson et al. (2005, page 336-339). The mean and variance of the distribution in (9) is given by

$$E(Y) = \frac{\mu}{1-\phi\mu}, \quad V(Y) = \frac{\mu}{(1-\phi\mu)^3} \tag{10}$$

The zero inflated generalized Poisson distribution has the probability mass function.

$$\begin{aligned} f(Y = y; P, \mu, \phi) &= P + (1-P)e^{-\mu} & y = 0 \\ &= (1-P)\frac{\mu(\mu + \phi\mu y)^{y-1} e^{-\mu-\phi\mu y}}{y!} & y = 1, 2... \end{aligned} \tag{11}$$

The mean and variance of zero inflated generalized Poisson distribution is given by

$$E(Y) = \frac{(1-P)\mu}{1-\phi\mu}, \quad V(Y) = \frac{(1-P)\mu}{1+\phi\mu}\left[\frac{P\mu}{1-\phi\mu} + \frac{1}{(1-\phi\mu)^2}\right]. \tag{12}$$

As in negative binomial and zero inflated negative binomial regression,here parameter μ of generalized poisson and Zero inflated generalized poisson regression is regressed to set of covariates $x_0, x_1, ...x_n$ as in (8)

3. Data description and model selection

This data is a part of record base cross sectional study conducted by Thiyagu (2010) for finding factors associated with adverse drug reaction for subjects admitted to the medicine department of a tertiary care hospital of Kasturba medical college, Manipal, India. The objective and exact description of the experiment is available in Thiyagu (2010). For present investigation, we have taken part of the data. The response variable is the number of adverse drug reaction. The factors considered are age, gender, type of system involved which was classified as central and peripheral nervous system, metabolic and nutritional system, gastrointestinal, skin and appendages, liver and bilateral system, type of drug which was classified as corticosteroids, anti asthmatic, anti tubercular, antibiotics, analgesics, diuretics and anti hypertensive.We had used SAS 9.2 version licensed to Manipal University for the present analysis.

The variables under consideration are summarized in Table 1. There were 445 subjects with 129 (29.9%)of people not having adverse drug reactions.Initially Poisson, negative binomial, generalized Poisson, Zero inflated Poisson, Zero inflated negative binomial and zero inflated generalized Poisson models were fitted to the data without covariates. BIC was used to assess goodness of fit of each model fitted in above step. Table 2 present the values of BIC and the probability of zero for each of the model. From the table it is clear that BIC ranges from 1642 (ZINB) to 1793 (Poisson). The percentage of zero count ranges from 0.15 (Generalized Poisson) to 0.29 (ZINB). The percentage of zero counts for the data is 29.9%. Therefore it was decided to select zero inflated negative binomial distribution as a suitable model for the data on adverse drug reaction.

For building a parsimonious model and to identify the risk factors for the number of adverse drug reaction in the second stage, covariates are included in the regression model based on zero inflated negative binomial distribution. Backward elimination procedure was used to identify the factors associated with adverse drug reaction. Table 3 summarizes the elimination procedure. From medical perspective, it is necessary that any selected model should include age, type of system involved as mandatory variables and thus are not included for the elimination. Initially 8 covariates are included in the model. The elimination procedure is carried out by deleting the covariates one by one and a model having minimum BIC (among the regression model with 7 covariates) is selected in the next step. The procedure of elimination of one variable at each step is carried out until there is no further change in BIC values. We stopped the procedure when BIC values started increasing by deleting any of the variables selected in the previous step. Thus, the model finally selected include four variables namely gender, anti asthmatic drugs, anti tubercular drugs and corticosteroids in addition to age and type of system involved. The estimated regression coefficient with standard error and 95% confidence interval of the final model are presented in Table 4. From the Table, it is clear that the four additional risk factors selected in the model are significant at 5% level of significance. Among the mandatory risk factors, age, type of system involved gastrointestinal turns out to be significant. Confidence interval is used to test for the significant of regression coefficients (Lehmann and Romano,2005). A regression coefficient is considered as significant if the corresponding confidence interval include zero as an interior point.

4. Binary logistic regression

Thiyagu (2010) has analysed the same data by converting the number of adverse drug reaction into binary. Since all variables considered in that study were not included for the present investigation, the data was reanalysed using logistic regression. Binary logistic regression was used to find factors associated with adverse drug reaction using backward elimination procedure. The outcome was classified as zero, if there is no ADR and one, if the is at least one ADR. Table 5 presents the elimination procedure for logistic regression. The selection criteria is same as in the case of count regression. Table 6 presents the details of the finally selected logistic regression model. Finally selected model includes three additional variables in addition to two mandatory variables. The selected variables are corticosteroids, anti tubercular and anti asthmatic drugs. All the three variables included in the model are significant at 5 % level of significance. As in the case of count regression, age, type of system gastrointestinal turned out to be a significant among mandatory variables.

5. Choice between the models

Regression models based on ZINB distribution included four additional variables while in the logistic regression it was only three. The risk factor which turn out to be significant in count regression model and did not appear in the logistic regression model is gender. The BIC for the ZINB regression model was 1610 while it was 1731 for logistic regression model. The difference in BIC values for these two models is 121 and is fairly a large difference. Table 7 presents the cross classification of the type of disease by gender of the subjects. From the table it is clear that the percentage of adverse drug reaction is significantly different for the two gender. This is confirmed by the Chi-square test with a P value of 0.018. When the above table is reduced to 2×2 table by treating one or more adverse drug reaction as a single entity, the Chi-square test turned out to be not significant. Gender has turned out to be a significant risk factor in previous studies on adverse drug reaction. For more details, see Nicolson (2010) and reference cited there in. Exponent of regression parameters in ZINB regression model will be rate ratio as compared to odds ratio in logistic regression. Our results are in agreement with McNutt et al. (2003) and Hilbe (2007), where they suggest use of count regression models than logistic regression for analysis of count data.

6. Conclusion

This work explores the consequences of converting a count response variable as binary variable.In present study, we had considered different types of count regression models namely Poisson,NB,ZINB, GP and ZIGP. For a count data set of adverse drug reactions, initially we fitted all these models and found that ZINB model fits well to the data.In second stage, we had included the covariates and fit the ZINB regression model.The same count dependent variable was categorized as binary and a logistic regression was fitted. We had found that ZINB regression model could able to include one more variable as compared to logistic regression model.Further the interpretation of regression coefficients is made in terms of rate ratio as compared to odds ratio in logistic regression models.Logistic regression is simple to compute and many softwares provide routines for building logistic regression models. Logistic regression is simple to interpret and is favorite choice among medical practitioners and researchers.The present investigation reveals that logistic regression may fail to identify the relevant risk factors as compared to ZINB regression models. Also ZINB model has an added advantage of modeling over dispersion in count data as compared to Poisson,ZIP regression models. However, since this investigation is confined to one case study further research is needed to underline this point.

References

[1] Claeskens, G. and Hjort, N. L. (2008): *Model selection and Model averaging.*Cambridge: Cambridge University Press.

[2] Consul, P. C. and Famoye, F. (1992): Generalized Poisson regression model. *Communications in statistics-Theory and Methods* 21, 89-109.

[3] Consul, P. C. and Shoukri, M. M.(1985): The generalized Poisson distribution when the sample mean is larger than the sample variance. *Communications in Statistics-Simulation and Computation* 14, 1533-1547.

[4] Consul, P. C. and Jain, G. C. (1973): A generalization of the Poisson distribution. *Technometrics* 15, 791-799.

[5] Cortina, L. M.(2005): Model Selection. *Encyclopaedia of Statistics in Behavioural Science* 3, 1251-1253, Chichester: Jhon Wiley and Sons.

[6] Cox, D. R. and Hinkley, D. V. (1979): *Theoretical statistics* (2nd edition). London: Chapman and Hall publication.

[7] Ghosh, J. K., Delampady, M. and Samanta, T. (2006): *An introduction to Bayesian Analysis: Theory and Methods.* USA: Springer texts in Statistics.

[8] Greenland, S. (2004): Model-based Estimation of Relative Risks and Other Epidemiologic Measures in Studies of Common Outcomes and in Case-Control Studies. *American Journal of Epidemiology* 60, 301-305.

[9] Hilbe, J. M. (2007): *Negative Binomial Regression.* Cambridge: Cambridge University Press.

[10] Hosmer, W. D. and Lemshow, S. (2000): *Applied Logistic Regression* (2nd edition). Canada: Wiley Publications.

[11] Jhonson, N. L., Kemp, A. W. and Kotz, S. (2005): *Univariate Discrete Distributions* (3rd edition). New Jersey: Jhon Wiley and Sons.

[12] Lehmann, E. L. and Romano, J. P. (2005): *Testing statistical hypotheses.* New York: Springer Science+ Business Media.

[13] McNutt, A. L., Wu, C., Xue, X. 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.

[14] Nelder, J. A. and Wedderburn, R. W. M. (1972): Generalized Linear Models. *Journal of Royal Statistical Society (Series A)* 135(3), 370-384.

[15] Nicolson, T. J., Mellor, H. R. and Roberts, R. R. (2010): Gender differences in drug toxicity. *Trends Pharmacol Sci* 31(3), 108-114.

[16] Schwarz, G. E. (1978): Estimating the dimension of a model. *Annals of Statistics* 6 (2), 461-464.

[17] Thiyagu, R. (2010): *Modeling of predictors for adverse drug reactions and pharmacoeconomic impact in a tertiary care hospital.* PhD [Thesis], Manipal University.

Table 1: Summary of characteristics of subjects assessed for adverse drug reaction

Variables (N=445)	n(%)
Number of adverse drug reactions	
0	129(28.9)
1	114(25.61)
2	80(17.97)
3	45(10.11)
4	24(5.39)
≥ 5	53(11.9)
Gender- Male	228(51.24)
Type of system involved	
Gastrointestinal system	62(13.93)
Central and pheripheral nervous system	96(21.57)
Skin and appendages	49(11.01)
Metabolic and nutritional	48(10.79)
Liver and Bilateral system disorders	37(8.31)
Type of drug given	
Diuretics	42(9.44)
Corticosteriods	21(4.72)
Anti tubercular	63(14.16)
Antibiotics	76(17.08)
Anti asthmatic	37(8.31)
Analgesics	20(4.49)
Anti hypertensives	33(7.42)

Table 2: Estimated inflate parameter and BIC values for different count regression models

Model	Prob. of Zero(Obs.=0.29)	BIC
Negative Binomial	0.26	1656
Poisson	0.15	1793
Generalized Poisson	0.15	1650
Zero inflated Poisson	0.28	1713
Zero inflated negative binomial	0.29	1642
Zero inflated generalized Poisson	0.16	1675

Table 3: Variables deleted in the backward elimination procedure for ZINB regression

Step	Variable eliminated	BIC
1	Antibiotics	1651
2	Antihypertensive	1643
3	Diuretics	1621
4	Analgesics	1610

Table 4: Estimated regression coefficients for ZINB regression

Variables	β	SE(β)	t ratio	P value	95% CI	
					lower	upper
Age	0.0089	0.0033	2.6837	0.0036	0.0024	0.0154
Gender	-0.1490	0.0700	-2.1286	0.0166	-0.2862	-0.0118
Antituberculotic drugs	0.5209	0.0226	23.0987	P<0.001	0.4767	0.5651
Anti asthmatic drugs	0.6677	0.3342	1.9980	0.0229	0.0127	1.3227
Cortico steriods	0.235	0.0529	4.4423	$P < 0.001$	0.1313	0.3386
Central & peripheral nervous system	0.3760	0.3342	1.1251	0.1303	-0.2790	1.0310
Metabolic& nutritional	0.0703	0.2587	0.2718	0.3929	-0.4367	0.5773
Gastro intestinal	-0.4995	0.1237	4.038	$P < 0.001$	-0.7415	-0.2570
Skin and appendages	-0.1667	0.2500	-0.6668	0.2524	-0.6567	0.3233
Inflate parameter P	0.1370	0.0332	4.1311	P<0.001	0.0720	0.2020
Dispersion parameter	0.2700	0.0944	2.8605	0.0021	0.0850	0.4550

Table 5: Variables deleted in the backward elimination procedure for logistic regression

Step	Variable eliminated	BIC
1	Diuretics	1821
2	Antihypertensive	1797
3	Analgesics	1762
4	Antibiotics	1747
5	Gender	1731

Table 6: Estimated regression coefficient for logistic regression

Variables	β	SE(β)	t ratio	P value	95% CI	
					lower	upper
Age	-.0028	.0014	-2.0149	.0220	-.0056	-.0001
Antituberculotic drugs	.5048	.0192	26.3448	P <0.001	.4672	.5423
Anti asthmatic drugs	.6291	.1294	4.86	$P < 0.001$	.3754	.8827
Cortico steriods	.189	.0429	4.4055	$P < 0.001$	.1049	.2730
Central & peripheral nervous system	.3429	.3318	1.0332	.1507	-.3075	.9933
Metabolic& nutritional	.0498	.2575	.1935	.4233	-.4549	.5545
Gastro intestinal	-.5385	.2697	-1.9963	.0230	-1.0672	-.0098
Skin and appendages	-.1782	.2485	-.7170	.2367	-.6653	.3089

Table 7: Cross classification of gender with number of ADR

Gender/No of ADR	0	1	2	3	4	$\geq$5	Total
Male	60	71	41	20	16	20	228
Female	69	43	39	25	8	33	217
Total	129	114	80	45	24	53	445

[1)]Chi-square=13.7025, df=6, P=0.018