Proceeding of ICSA 2019, p: 1-8
ISBN 978-979-19256-3-1 (PDF)

Comparison of Poisson and Quasi-Poisson Regression: A Simulation study

A Gabriella[1], S Abdullah[2*], S M Soemartojo[3].

[*1,2,3)]*Department of Mathematics, Universitas Indonesia, Kampus Baru UI, Depok, 16424, Indonesia*

E-mail: sarini@sci.ui.ac.id

Abstract

Poisson regression is often used to model count data. However, it requires the assumption of equidispersion which not always met in the real application data. Quasi-Poisson can be considered as an alternative to handle this problem. The objective of this essay is to explain about the Quasi-Poisson regression, the likelihood construction, parameter estimation, and its implementation in real life data. The numerical method used in this study is Newton-Raphson which is equivalent to Iterative Weighted Least Square (IWLS) at the end of calculation. The simulation results for the data with the above problem showed that, in case of overdispersion, Quasi-Poisson regression with Maximum Quasi-Likelihood method provided a good fit to the data compared to Poisson regression.

***Keywords*:** Iterative weighted least square, Newton Raphson, overdispersion, quasi-likelihood, quasi-Poisson.

Introduction

Poisson regression is often used when the response variable contains count data. In application, Poisson regression has several assumptions that must be fulfilled. One of the Poisson regression assumption is equidispersion where mean response variable and its variance are equal. However, in practice, overdispersion is more common than equidispersion (McCullagh and Nelder, 1989). Unobserved heterogeneity is one of the cases that leads to overdispersion, where variance exceed the mean (Cameron and Trivedi, 1998). Overdispersion data certainly requires special handling to be analyzed.

Regression model that can handle overdispersion problem is Quasi-Poisson regression model (Ver Hoef and Boveng, 2007), where this model pays attention to the dispersion parameter which causes the data variance unequal to the mean. This model starts from parameter estimation called quasi-likelihood. In likelihood estimation, response variable should follow some distribution while quasi-likelihood estimation only requires the relationship between mean and variance from response variable. Moreover, quasi-Poisson regression model estimate the value of dispersion parameter while Poisson regression did not pay attention to the value of dispersion parameter.

In terms of regression parameter estimation, quasi-Poisson model is very related to quasi-likelihood. Quasi-likelihood function, which is the core of this study, has the same uses as likelihood function in maximum likelihood estimation (Wedderburn, 1974). However, there are prominent difference between this two parameters estimation that will be explained in the next section. Numerical method used for regression parameter estimation through quasi-likelihood is newton Raphson which equal to estimate the parameter with Iterative Weighted Least Square (IWLS) (Ver Hoef and Boveng, 2007).

Data simulation will be done in terms of showing that quasi-Poisson model is better than Poisson model when the response variable is over dispersed count data. Regressor will be set

in two variables where the first regressor (x_1) represent the significant variable and the second regressor (x_2) represent the variable that not significant in modelling response variable. The value of standard error of regression parameter from quasi-Poisson regression will exceed the value of standard error of regression parameter from Poisson regression because of the parameter dispersion effect and it will affect the determination of significant variables. Data simulation is completed using the help of software RStudio 3.5.3 version (RStudio Team, 2015).

1. Materials and Methods

Quasi-Poisson model formed by generalized linear models with Poisson-like assumption (Ver Hoef and Boveng, 2007). The response variable probability density function (pdf) in quasi-Poisson model must be included in one parameter exponential family. Assumption of response variable is written as follows

$$E(Y) = \mu,$$
$$Var(Y) = \phi\mu,$$

where μ denote the mean response variable Y and ϕ denote the dispersion parameter which will be estimate from the data. This paper will define dispersion parameter value greater than one to state the overdispersion condition.

Likelihood function for quasi-Poisson (quasi-likelihood) does not require a specific probability density function to estimate regression parameter except for response variable assumption (McCullagh and Nelder, 1989). Formation of quasi-likelihood function begins with the same way as the usual likelihood function with the general pdf from exponential family (McCullagh and Nelder, 1989). Let Y be the response variable is part of one parameter exponential family, likelihood function of Y is

$$L(\theta) = f(y_1, y_2, \dots, y_n|\boldsymbol{\theta}) = f(y_1|\boldsymbol{\theta})f(y_2|\boldsymbol{\theta}) \dots f(y_n|\boldsymbol{\theta})$$

$$= \exp\left[\sum_{i=1}^{n}\left(\frac{y_i\theta - B(\theta)}{\phi} + C(y_i,\phi)\right)\right],$$

then

$$\ln\big(L(\theta)\big) = \sum_{i=1}^{n}\left(\frac{y_i\theta - B(\theta)}{\phi} + C(y_i,\phi)\right).$$

From the exponential family, θ is canonical parameter which is a function of μ ($\theta = g(\mu)$) and μ is a function of β (Agresti, 2013). Regression parameter β is the parameter that will be estimated, therefore $\ln(L(\theta))$ will derive toward μ.

$$\frac{\partial \ln(L(\theta))}{\partial \mu} = \frac{\partial}{\partial \mu}\left[\sum_{i=1}^{n}\left(\frac{y_i\theta - B(\theta)}{\phi} + C(y_i,\phi)\right)\right]$$

$$= \frac{1}{\phi}\left[\sum_{i=1}^{n}\left(y_i\frac{\partial\theta}{\partial\mu} - \frac{\partial B(\theta)}{\partial\mu}\right)\right],$$

where $B(\theta)$ denote a function of θ and θ is a function of μ. Consequently, chain rule will be used.

$$= \frac{1}{\phi}\left[\sum_{i=1}^{n}\left(y_i\frac{\partial\theta}{\partial\mu} - \frac{\partial B(\theta)}{\partial\theta}\frac{\partial\theta}{\partial\mu}\right)\right]$$

$$= \frac{1}{\phi}\left[\sum_{i=1}^{n}\frac{\partial\theta}{\partial\mu}(y_i - B'(\theta))\right].$$

Based on exponential family, $B'(\theta) = \mu$ and $B''(\theta)\phi = V(\mu) = Var(Y)$.

$$= \sum_{i=1}^{n}\frac{(y_i - B'(\theta))}{\frac{\partial B'(\theta)}{\partial\theta}\phi}$$

$$= \sum_{i=1}^{n}\frac{(y_i - \mu)}{B''(\theta)\phi}$$

$$\frac{\partial \ln(L(\theta))}{\partial \mu} = \sum_{i=1}^{n}\frac{(y_i - \mu)}{V(\mu)}.$$

It is shown that to obtain derivative of log likelihood for one parameter exponential family response variable toward μ it only requires the relationship between mean and variance from response variable (Wedderburn, 1974). Hence, the form that used in parameter estimation without paying attention a specific pdf of variable called quasi-likelihood (Wedderburn, 1974). The following function represent the quasi-likelihood function for quasi-Poisson denote with $Q(.,.)$ (Wedderburn, 1974)

$$\frac{\partial Q(y_i,\mu_i)}{\partial \mu_i} = \frac{y_i-\mu_i}{\phi \mu_i},$$

$$Q(y_i,\mu_i) = \int_{y_i}^{\mu_i} \left(\frac{y_i-\mu_i^*}{\phi \mu_i^*}\right) \partial \mu_i^*.$$

Quasi-Poisson regression model is one of generalized linear model with link function log or log link which written as follows

$$\ln(\mu) = \mathbf{x^T}\boldsymbol{\beta} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

where μ represent mean response variable, x_j denote regressor or predictor variable with $j = 1,2,\ldots,p$, and β_j denote regression parameter.

Regression parameters from quasi-Poisson can be calculated using iterative weighted least square (IWLS) (Ver Hoef and Boveng, 2007). The equation used in calculated the iterative weighted least square with $k+1$ iteration is written bellow (Ver Hoef and Boveng, 2007)

$$\widehat{\boldsymbol{\beta}}^{[\boldsymbol{k+1}]} = \left(\mathbf{X^T W^{[k]} X}\right)^{-1} \mathbf{X^T W^{[k]}} \tilde{\mathbf{y}}^{[\boldsymbol{k}]}$$

where

$$\tilde{y}_i^{[k]} = \eta_i^{[k]} + (y_i - \mu_i^{[k]}) \frac{\partial \eta_i^{[k]}}{\partial \mu_i^{[k]}}, \qquad \eta_i^{[k]} = \mathbf{x_i^T}\boldsymbol{\beta}^{[k]}, \qquad \mu_i^{[k]} = g^{-1}(\eta_i^{[k]}),$$

and $\mathbf{W}^{[k]}$ denote the weight matrix which is diagonal matrix with diagonal elements

$$w_i^{[k]} = \frac{1}{\mu_i^{[k]}} \left(\frac{\partial \mu_i^{[k]}}{\partial \eta_i^{[k]}}\right)^2.$$

Simulation design

Data used in simulation are from generated data with uniform distribution using R studio with *glm* package which will explained in next chapter. Then the data will be analyzed with quasi-Poisson regression that already explained in previous chapter and Poisson regression model. The result from both model will be compared to obtain conclusion.

In this paper, quasi-Poisson simulation aims to show that quasi-Poisson modelling overdispersed data better than Poisson regression. Data for explanatory variables, x_1 and x_2, were generated from uniform distribution, regression coefficients were set to $\beta_0 = 1.2,\ \beta_1 = 1.7,\ \beta_2 = 0$ to accommodate that x_1 as a significant variable and x_2 as nonsignificant in terms of modelling y. Then the response variable, Y is obtained as $Y \sim Poi(\lambda; d)$ where $\log(\lambda) = \boldsymbol{X\beta}$ and $d = 3$ denote dispersion parameter. The last step is modelling with quasi-Poisson and Poisson regression then compare the standard error and p-value from both regressions.

Results

The fitted regression from the simulation is $\ln(\mu) = 1.2 + 1.7x_1 + (7.804 \times 10^{-11})x_2$, where the dispersion parameter for quasi-Poisson and Poisson are 3.032102 and 1 based on the results of the program.

Based on the result of the data simulation, there are prominent difference in the dispersion parameter comparing both regressions. The differences between quasi-Poisson and Poisson are presented in Table 1 that contains standard error and p-value from both regressions.

Table 1. Standard errors of quasi-Poisson regression and Poisson regression in simulation.

Parameter	Standard Error		p-value	
	Quasi-Poisson	Poisson	Quasi-Poisson	Poisson
Intercept	1.341e-08	7.702e-09	<2e-16	<2e-16

β_1	6.022e-10	3.458e-10	<2e-16	<2e-16
β_2	6.060e-11	3.480e-11	0.198	<0.0249

Table 1 shows that quasi-Poisson regression was able to model overdispersed data better than Poisson regression because quasi-Poisson consider x_2 as a not significant (p-value > 0.05) variable while Poisson consider x_2 as a significant variable (p-value < 0.05).

Conclusion

The simulation showed that quasi-Poisson regression fit the data better than Poisson regression in the case of over-dispersed data. Result and simulation conclude that quasi-Poisson regression model can overcome overdispersion problem with the appearance of dispersion parameter.

Acknowledgements

This research supported by the University of Indonesia with PITTA B 2019 research grant scheme, with ID number NKB-0665/UN2.R3.1/HKP.05.00/2019. We thank to all reviewers for the improvement of this article.

References

Agresti, A. 2013. *Categorical Data Analysis* (3 ed.). New York: John Wiley & Sons.

Cameron, A. C., & P. K. Trivedi. 1998. *Regression Analysis of Count Data.* United Kingdom: Cambridge University Press.

McCullagh, P., and J. A. Nelder. 1989. *Generalized Linear Models* (2 ed.). New York: Chapman & Hall.

RStudio Team (2015). RStudio: Integrated Development for R. RStudio, Inc., Boston, MA URL http://www.rstudio.com/.

Ver Hoef, J.M. and Boveng, P.L. 2007. Quasi-Poisson Vs. Negative Binomial Regression: How Should We Model Overdispersed Count Data?. *Ecology*, Vol. 88, No. 11.

Wedderburn, R.W. M. 1974. Quasi-Likelihood Functions, Generalized Linear Models, and The Gauss-Newton Method. *Biometrika*, Vol. 61, No. 3.