Available online at www.asric.org
ASRIC Journal on Natural Sciences 1 (2021) 30-37

Scaling Outcome Correlated Binary Big Data Using Estimates of Bivariate Dispersion Parameters

Ahmed M. M. Elsayed [1], Nevein N. Aneis[1]

[1]Al-Obour High Institute For Management & Informatics, Department of Basic Sciences, Kilo 21 Cairo-Belbies Road, P.O. Box 27 Al-Obour City, Egypt

Corresponding Author Email: atabl@oi.edu.eg.

Received 31 May 2021; revised 13 July 2021; accepted 23 August 2021

***Abstract*— Dispersion parameter should be the unity in the case of the univariate Bernoulli data. But there may be some deviations if there is a sequence of the Bernoulli outcomes, that may lead to Binomial case. Over (lower) dispersion criterion is happened if the variance of actual response, var(y), is more (less) than the nominal variance as a function of the mean, var(μ). This paper presents the mathematical form for estimating and modifying the dispersion parameters for the outcome correlated binary (0,1) Big data, with scalar and matrix values, in Bivariate case. The impact of the estimates of dispersion parameter on the outcome correlated binary Big data is indicated.**

In general, the aim is making the dispersion parameters are close or equal to the unity. The purpose is controlling of marginal probabilities of the correlated binary outcomes. Since the increasing of marginals, increases the values of dispersion estimates. We can use these property to decrease the over-dispersion to close to the unity. The R program and its packages, is used to generate and fit the binary correlated Big data. Scaling and Roots techniques that depend on the estimates of dispersion parameters are used to modify the outcome correlated binary data. We have found that Scaling and Roots processes have similar results and good effects, only for binary Big data. Since the manner is different when deal with Small observations.

***Keywords*— VGAM, VGLM, Binary outcomes, Dispersion parameters, Big data, Scaling data, Correlated data.**

I. INTRODUCTION

The estimation of dispersion parameter in the univariate case can be obtained easily using the Pearson's Chi-square or the Deviance function. The over(lower) dispersion can be deduct from the equation: $\text{var}(y) = \phi \ \text{var}(\mu)$ where ϕ is the univariate dispersion parameter. When $\phi > 1$ this implies the over-dispersion, while $\phi < 1$ implies the lower-dispersion (McCullagh and Nelder, 1989). Many studies have devoted the dispersion criteria in Univariate case, namely, when the Binomial data are used. It is difficult to extend these methods to estimate the dispersion parameters in Bivariate case. Because in Bivariate case, the association between the correlated response variables may be happened. So, we must take this association into account when estimate the dispersion parameter. In Independence case, the estimate of dispersion parameter ϕ is performed as in the univariate case.

Some studies have presented attributes of the overdispersion problem as Smith and Heitjan (1993) provided an appropriate statistical tool to detect extra Binomial variation.

Cook and Ng (1997) described Bivariate logistic-normal mixture model for over-dispersed two state Markov processes. Saefuddin et al. (2011) showed the effect of overdispersion on the hypothesis test of Logistic regression. Simple method proposed by William (1982) was used to correct the effect of overdispersion by taking the inflation factor into consideration. When the overdispersion does not occur or very small overdispersion occurs, dispersion parameter ϕ will be approximately equal to zero, so Y_i exactly follows Binomial distribution, $Bin(n_i, \pi_i)$, and $Var(Y_i) = n_i\pi_i(1-\pi_i)$, Collett (2003). The value of Pearson's Chi-square statistic depends on $\hat{\phi}$ so, iteration process, is needed to find the optimum value, as a test of $\hat{\phi}$.

Dispersion parameter for binomialff family =1. For the two correlated binary Big data, the independent variable X_3 has the lowest residual deviance, this reflects the importance of this variable to the model. Also, has a significant effect with 2nd additive predictor in the case of 500,000 observations. But there are not one of the other independent variables X_1 , X_2 have significant effects, with 5% significant level. For Loglikelihood value, the independent variable X_1 has the lowest value in the case of 10,000 observations. While the independent variable X_3 has the lowest value in the case of 500,000 observations.

V. CONCLUSION

This paper presents the mathematical form for estimating and modifying the dispersion parameters for the outcome correlated binary (0,1) Big data, with scalar and matrix values, in Bivariate case. The effect of dispersion estimates on the outcome correlated binary Big data is indicated. The marginals of two correlated binary outcomes variables effect on the values of estimates of dispersion parameters. Using these property, we can motivate the tends of estimates to close to the unity. The program R and its packages are used to generate and fit the Big data.

Roots and Scaling methods are used to modify the outcome correlated binary Big data. We have found that Scaling and Roots processes have similar results, and good effects only for binary Big data. Since the manner is different when deal with small obsevations.

ACKNOWLEDGMENTS:

For Al-Obour High Institute For Management & Informatics.

REFERENCES

[1] McCullagh, P. and Nelder, J. (1989). Generalized linear models (second edition), Chapman and Hall, London, United Kingdom.

[2] Smith, P. and Heitjan, F. (1993). Testing and adjusting for departures from nominal dispersion in generalized linear models, Applied Statistics 42, 1 : pp 31–34.

[3] Cook R. and Ng, E. (1997). A logistic-bivariate normal model for overdispersed two-state Markov process, Biometrics 53, 1 : 358-364.

[4] Saefuddin, A. Setiabudi, A. and Achsani, N. (2011). The effect of overdispersion on regression based decision with application to Churn Analysis on Indonesian Mobile Phone Industry, European Journal of Scientific Research 60, 4 : 584-592.

[5] William, D. (1982). Extra-binomial variation in logistic linear models, Applied Statistics 31 : 144-148.

[6] Collett, D. (2003). Modeling binary data (Second edition), Chapman and Hall, London, United Kingdom.

[7] Davila, E., Lopez L. and Dias, G. (2012). A statistical model for analyzing interdependent complex of plant pathogens, Revista Colombiana de Estadistica Numero especial en Bioestadistica 35, 2 : 255-270.

[8] Casella, G. and Berger , R. (2002). Statistical inference (second edition), Duxbury Press, Florida, United States.

[9] Hurvich, C. and Tsai, C. (1989). Regression and time series model selection in small samples, Biometrika, 76 : 297–307.

[10] Yee, T. and Wild, C. (1996). Vector generalized additive models. Journal of the Royal Statistical Society, Series B, Methodological, 58 : 481–493.

[11] Yee, T. and Hastie, T. (2003). Reduced-rank vector generalized linear models. Statistical Modelling, 3 :15–41.

[12] Yee, T. (2004). A new technique for maximum-likelihood canonical Gaussian ordination. Ecological Monographs, 74 : 685–701.

[13] Yee, T. (2006). Constrained additive ordination. Ecology, 87 : 203–213.

[14] Yee, T. and Stephenson, A. (2007). Vector generalized linear and additive extreme value models. Extremes, 10 : 1–19.

[15] Yee, T. (2008). The VGAM Package. R News, 8 : 28–39.

[16] Yee, T. (2010). The VGAM package for categorical data analysis. Journal of Statistical Software, 32 : 1–34.

[17] Yee, T. (2014). Reduced-rank vector generalized linear models with two linear predictors. Computational Statistics and Data Analysis, 71 : 889–902.

[18] Yee, T. (2015). Vector Generalized Linear and Additive Models: With an Implementation in R. New York, USA: Springer.

[19] Elsayed, A. Islam, M. and Alzaid, A. (2013). Estimation and test of measures of association for correlated binary data, Bulletin of the Malaysian Mathematical Sciences Society 2, 36, 4 : 985-1008.

[20] Elsayed, A. (2016). A new approach for dispersion parameters. Journal of Applied Mathematics and Physics (JAMP) 4, 8 : 1554-1566.