MATEMATIKA, 2013, Volume 29, Number 1b, 17-23

An Application of Logistic Regression on Correlated Data

Asep Saefuddin
Department of Statistics, Faculty of Mathematics and Natural Sciences, Bogor Agricultural University
Jl. Raya Darmaga Kampus IPB Darmaga Bogor
16680 West Java, Indonesia
e-mail: asaefuddin@gmail.com

Abstract It is common to find data in the form of proportions have more variability than the variance based on the binomial distribution. The phenomenon is called extra-binomial variation or overdispersion and commonly caused by the occurrence of correlation within response variable. Models for correlated outcome produces unbiased estimate, but its standard error is underestimated. Hence, confidence interval becomes too narrow and statistical test tends to reject the null hypothesis. Williams has proposed to simple method correct the effect of extra-binomial variation by taking inflation factor into consideration. In this paper, the Williams approach is implemented to analyze the poverty data in Indonesia, which exhibit extra-variation. The result shows that the method adjusts the standard error of estimates and then provides more reliable conclusion than the naive approach. Public policy of government certainly requires adequate recommendations to allocate limited resources appropriately following defined objectives. Regional data usually depends on many factors causing non-independent outcome making the data are overdispersed. Models which ignore the extra-variation will lead to a wrong conclusion. Therefore, applying regression analysis in public policy with accommodating overdispersion is important to obtain meaningful and reliable recommendations.

Keywords Logistic Regression; Extra-binomial Variation; Overdispersion; Williams Method; Poverty Analysis.

2010 Mathematics Subject Classification 62J05

1 Introduction

It is common to find that the data is grouped in the form of proportion or count producing binomial or Poisson distribution. When we have 'n' observation with 'p' is the probability of 'success', the binomial variance is assumed to be constant and equal to $np(1-p)$. It implies that the individual binary outcome is independent one to another. In epidemiologic study, independent responses are uncommon which then produce the variance is greater than that on the assumption of binomial variability. This phenomenon is called extra binomial variation or overdispersion, [1]. On the other hand, epidemiology usually applies logistic regression to analyze the effect of risk factors on the response probability. When the data exhibit overdispersion, results from the logistic models may not be appropriate. Hence, correction action is needed.

The case of overdispersion does not only occur in the epidemiologic study, but also in the other areas of observational studies. Socioeconomic data, for example, are not easy to find independent outcome variables. Many regional factors affect the variable of interest which then produces overdispersion of the outcome. The geoinformatics group of the Department of Statistics of Bogor Agricultural University has compiled statistical methods for socioeconomics data and relaxed some rigid assumption. The complexed socioeconomic problems in Indonesia need to be solved by tailor-made comprehensive analysis using

advanced statistical approaches. While, the simplest one is to handle extra binomial variation using Williams method for correlated data.

This paper presents a simple approach for handling correlated outcome using Williams approach. The example is taken from poverty data in the form of proportion which then likely exhibits overdispersion due to intraclass correlation. It is understandable since poverty is usually influenced by many factors within locations, including the local government policy and natural diversity (agro-ecological typology). Angraini, [2] has proposed to handle the intraclass correlation and the spatial dependency simultaneously.

2 Methods: Brief Theoretical Review

2.1 Overdispersion

Overdispersion in a simple definition is an extra variation in which the real variance exceeds the variance based on its assumed distribution. In the case of binomial theory it is termed the extra binomial distribution. In binomial regression, there are many factors causing residual variance significantly larger than its expectation indicated by the residual mean deviance is greater than one. It may be caused by inadequate models for non-independent data. Some advanced approaches to tackle the overdispersion are inclusion random effect in the models, applying the beta-binomial distribution, and weighted regression. This paper applies the simple approach proposed by [1], i.e. iterative weighted regression. The approach is very simple but useful, especially when the overdispersion is caused by intraclass correlation yielding non-independent grouped response variable.

2.2 Binary Logistic Regression

Suppose there are k binomial observations written in the form of proportion 'y_i/n_i' where y_i is number of occurrences of event and n_i is number of population within i^{th} observation (i = 1, 2 , ..., k). Consider that π_i is probability of event and $E(Y_i)=\pi_i n_i$ is expected value for each random variable with variance $\text{var}(Y_i)=\pi_i n_i(1-\pi_i)$. Logistic regression model for π_i with p predictors (X_1, X_2, …, X_p) expressed as logit function of event is

$$\text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_p X_{pi} \tag{1}$$

The expression is known as Generalized Linear Model (GLM) with logit link function, [3], [4]. Parameters are estimated using maximum likelihood method with the likelihood function is

$$L(\beta) = \prod_{i=1}^{k} \binom{n_i}{y_i} \pi_i^{y_i} (1-\pi_i)^{n_i - y_i} \tag{2}$$

Obtaining parameter estimate, $\hat{\beta}$, so that $L(\hat{\beta})$ or natural log of $L(\hat{\beta})$ reach maximum value using iteration process, such as by Newton-Raphson method, ([5], [6]) and Fisher's scoring ([4], [6], [7]). For many GLMs, including binary models of logit link, with full-rank Hessian matrix is negative definite and the log likelihood is a strictly concave function. Therefore maximum likelihood estimates of model parameters exist and are unique under quite general conditions, [5].

2.3 Deviance Statistic and Pearson's Chi-square

Deviance statistic and Pearson's chi-square measure lack-of-fit, and goodness-of-fit at once, of logistic regression model. If fitted value of number of 'event' is denoted by $\hat{y}_i$, where $\hat{y}_i = \hat{\pi}_i n_i$, its deviance statistics (D) is

$$D = 2\sum_{i=1}^{k}\left\{ y_i \log\left(\frac{y_i}{\hat{y}_i}\right) + \left(n_i - y_i\right) \log\left(\frac{n_i - y_i}{n_i - \hat{y}_i}\right)\right\} \tag{3}$$

whereas Pearson's chi-square statistic (X^2) of the model is obtained by

$$X^2 = \sum_{i=1}^{k} \frac{(y_i - \hat{\pi}_i n_i)^2}{\hat{\pi}_i n_i (1 - \hat{\pi}_i)} \tag{4}$$

Both D and X^2 statistic follow chi-square (χ^2) distribution with $k - p$ degree of freedom, where k is number of binomial observations and p is numbers of parameters included in the model. If logistic model is satisfactory, D and X^2 statistic will be close to its number of degree of freedom. Or in other word, ratio of D or X^2 over the degree of freedom will be close to unity, [7].

2.4 Modeling Overdispersion

When ratio of either D or X^2 significantly exceeds one, assumption of binomial variation is violated and then overdispersion is occurred. In reality, non-independent object with positive correlation will inflate variance of response probability and hence called overdispersion. The other case which is underdispersion may be caused by negative correlation among objects, [7].

Theoretically, overdispersion does not affect the expected value of binomial distribution and maximum likelihood estimated parameter is unbiased. However, its standard error will be underestimated due to overdispersion. Then, confidence interval of parameter will be narrow, and as a result, null hypothesis that parameter is equal to zero tend to be rejected. In other word, the explanatory variables tend to have significant effect to the response.

Modeling of overdispersion is often expressed in equation of variance of response variable, Y_i, as

$$\text{var}(Y_i) = \pi_i n_i (1 - \pi_i)\left\{1 + (n_i - 1)\phi\right\} \tag{5}$$

where $\{1 + (n_i - 1)\phi\}$ is overdispersion scale and ϕ denotes inflation factor. When overdispersion is not occurred or very small, ϕ will equal to or approximately zero, hence Y_i follows binomial distribution with mean $E(Y_i) = \pi_i n_i$ and variance $\text{var}(Y_i) = \pi_i n_i (1 - \pi_i)$. However, if overdispersion is occurred, ϕ exceeds zero and leads var(Y_i) to be greater than binomial variance, [7]. Hence, the empirical sampling variance is greater than the theoretical one.

2.5 Williams Method

A procedure for handling overdispersion on logistic regression has been firstly introduced by [1], and then has been called as Williams approach, [7]. It corrects the weights of estimating logistic regression parameter. Williams method uses the inverse of overdispersion scale to weight binomial observation rather than constant value. As on general linear model ([4]), this is to stabilize the variance between response probabilities in the logistic regression model when $\text{var}(Y_i) = \pi_i n_i (1 - \pi_i)$. Associated weights for $i^{\text{-th}}$ binomial observation using Williams method is expressed as

$$w_i = 1/[1 + (n_i - 1)\phi] \tag{6}$$

The main principle of the method is to obtain an optimum value of inflation factor to be put into Equation 7. Parameter estimate for inflation factor ϕ, or $\hat{\phi}$, is obtained by equating X^2 statistic of the logistic model to its expected value, written as:

$$X^2 = \sum_{i=1}^{k} \frac{w_i (y_i - \hat{\pi}_i n_i)^2}{\hat{\pi}_i n_i (1 - \hat{\pi}_i)} \tag{7}$$

and

$$E\left(X^2\right) \simeq \sum_{i=1}^{k} w_i (1 - w_i v_i d_i) \left\{1 + (n_i - 1)\hat{\phi}\right\}$$

where $v_i = \pi_i n_i (1 - \pi_i)$, w_i is the weight and d_i is diagonal element of the variance-covariance matrix of the linear predictor, $\hat{\eta}_i = \sum \hat{\beta}_j x_{ji}$. The value of X^2 statistic and $\hat{\phi}$ are dependent each other, hence iteration process is absolutely needed to find the optimum value. The algorithm of Williams method has been described deeply by [1] and also rewritten by [7] and [8].

Once the inflation factor estimate $\hat{\phi}$ has been optimized, w_i can be used in a weighted fit of model (Collett 2003). If $\hat{\phi} = 0$, standard logistic regression in Equation (1) is held. Note, Williams method can only be used for if response variable is in 'y_i/n_i' form, [6].

3 Application on Poverty Analysis

3.1 Data and Modeling

The model is on proportion of number of poverty (POV) over number of population (POP) as a function of Human Development Index (HDI), monthly adjusted per-capita expenditure in thousand rupiahs (EXPD), percentage of labour participation (LABR), per-capita Gross Regional Domestics Product at current market prices in thousand rupiahs (GRDP) and status of region (1 for city and 0 for regency). The data is provided by The National Team for Accelerating Poverty Reduction, The Office of Vice President of Republic of Indonesia (2010) for the year of 2008. Observation units are 116 regencies and cities in Java, Indonesia, spread out in six provinces. Due to the nature of poverty, the outcome variable (POV) as proportion may not follow the assumption of independent outcome.

Suppose there are 116 regions (k=116) taking place as binomial observations written in the form of proportion 'POV_i/POP_i' in the i^{th} region (i = 1, 2 , ..., 116). Consider that Pr_i is probability of poor people within i^{th} region. According to Equation (1), general form of logistic regression model for poverty can be expressed as

$$\log\left(\frac{Pr_i}{1 - Pr_i}\right) = \beta_0 + \beta_1 HDI_i + \beta_2 EXPD_i + \beta_3 LABR_i + \beta_4 GRDP_i + \beta_5 STAT_i \tag{8}$$

Standard logistic regression model in Equation (8) is firstly implemented, and then goodness-of-fit of the model is measured using deviance and Pearson's chi-square statistics. If these statistics turn out the occurrence of overdispersion, Williams methods is used to solve the problem. Furthermore, the results of standard logistic regression is compared to the results of Williams method.

3.2 Result of Standard Logistic Regression and Logistic Regression using Williams Method

Data are fitted to standard logistic regression model following the Equation (8). Analysis of maximum likelihood estimates of the model are summarized in Table 1. The p-values of Wald chi-square test, which test the null hypothesis that parameter equal to zero, are all smaller than 0.01. It means null hypothesis for each parameter is rejected at α=1%.

Table 1 Analysis of Maximum Likelihood Estimates of Standard Logistic Regression Model and Logistic Regression Model With Williams Method

Parameter	DF	Standard Logistic Regression Model				Logistic Regression Model With Williams Method			
		Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	1.18320	0.049700	566.0412	<.0001	4.97770	7.59700	0.4293	0.5123
HDI	1	-0.06680	0.000097	470014.3400	<.0001	-0.05430	0.01420	14.7091	0.0001
EXPD	1	0.00168	0.000086	383.8463	<.0001	-0.00569	0.01310	0.1886	0.6641
LABR	1	0.01260	0.000057	49025.0537	<.0001	0.01260	0.00814	2.3874	0.1223
GRDP	1	-6.12E-6	2.077E-8	86848.4970	<.0001	-3.02E-6	1.756E-6	2.9591	0.0854
STAT	1	-0.44050	0.001050	177154.1770	<.0001	-0.38700	0.12010	10.3880	0.0013

The standard regession assuming independent objects produces all predictors are highly significant affecting the poverty (α<1%). In this case, in order to reduce the poverty, all factors have to be considered with similar attention. If the government has unlimited resources, the action will be affordable, although cost inefficient. To have cost efficient and appropriate priority, a creative analytical approach is needed.

The value of deviance statistic is 2.3×10^6 with 110 degree of freedom and the value of Pearson's chi-square is 2.4×10^6 on 110 degree of freedom (Table 2). These indicate poverty data contain overdispersion, where the ratios of the statistics to their degrees of freedom are all significantly greater than one. This phenomenon seems reasonable since the socio-economic condition of people within region is likely affected by community. Therefore, the probability of people being poor is also correlated within the region termed as the intraclass effect.

Once overdispersion presents, logistic regression model violates the binomial assumption. Therefore, previous conclusion related on statistical test is not reliable and misleading, hence need a revision. The results from both approaches are listed in the Table 1.

Table 2 Deviance and Pearson Goodness-of-Fit Statistics of Standard Logistic Regression Model and Logistic Regression Model With Williams Method

Criterion	DF	Standard Logistic Regression Model			Logistic Regression Model With Williams Method		
		Value	Value/DF	Pr > ChiSq	Value	Value/DF	Pr > ChiSq
Deviance	110	2349469.39	21358.81	<.0001	107.7737	0.9798	0.5422
Pearson	110	2423555.10	22032.32	<.0001	109.9981	1.0000	0.4821

Since overdispersion is occurred, the logistic regression model is then adjusted using Williams method. Through iteration process, estimate of inflation factor is 0.019538, hence overdispersion scale is $[1+(POP_i-1)*0.019538]$. Using the weight of $w_i=1/[1+(POP_i-1)*0.019538]$ the calibrated logistic regression model is implemented for estimating the regression coefficients. The results are summarized in Table 1, while the goodness-of-fit of the model according to deviance and Pearson's chi-square statistic are listed in Table 2.

From Table 2, it is obvious that William method corrects the lack-of-fit of model indicated by reduction of values of deviance and Pearson's chi square statistic. Deviance statistic and Pearson's chi-square of logistic model with William method are 107.773 and 109.998, respectively, which is very close to the degree of freedom 110. These quantities indicate that overdispersion problem has been accommodated appropriately using William approach. Therefore, more reliable conclusion can be drawn from Table 1 of Williams method.

From the analysis's point of view, the poverty data is approached by accommodating the possibility of overdispersion which indeed is true. Table 1 indicates that only HDI (Human Development Index) and

STAT (dummy variable of regional status) provide significant effect on poverty. Therefore, the government may concentrate on these factors. In other word, to reduce poverty in Indonesia is focused just to improve factors related to HDI and regional status, i.e. education, health, and market facilities. The model also recommends a follow up actions related to the human quality of life. The model of poverty in this paper is just on inception analysis which needs further approach based on a more complex model.

4 Conclusion

Correlated object of interest may cause standard error of the estimates to be underestimated. The phenomenon is common in logistic regression to model proportion outcome. Models ignoring correlation among objects tends to reject null hypothesis or to accept the alternative one which actually may be false. In the binomial regression it is necessary to get insight on data generation such as indicated in poverty data. When the intraclass correlation exists, all parameter estimates yield significant effect to the outcome. It is actually because of correlation within observation rather than influence of predictors. Therefore, statistical recommendations, such as on how strategy for alleviating poverty should be campaigned, is slightly distorted. False analysis will not only bring inappropriate recommendation, but also causes wastage due to excessive cost allocation and resources. The Williams method is a simple approach in handling overdispersion, but it is not the only one. There are many methods could be applied, for example, beta-binomial regression ([9], [10] and [11]), mixed effect model ([12]), logistics normal model ([11]), quasi-likelihood method ([13]) and double-exponential method approach (*see* [14]). When the overdispersion is due to correlation among objects, the William approach is appropriate.

Acknowledgements

I am very grateful to the geoinformatics team of the Department of Statistics IPB who has dedicated doing research on applying statistics for poverty analysis in Indonesia. Also this acknowledgment is extended to my colleagues at UKP4-Jakarta.

References

[1] Williams, D.A. Extra-Binomial Variation in Logistic Linear Models. *Applied Statistics*. 1982. 31(2): 144-148.
[2] Angraini, Y. *Spatial Panel Data Analysis*. PhD research. 2012.
[3] McCullagh, P. and J.A. Nelder. *Generalized Linear Models, Second ed*. London, UK : Chapman and Hall. 1989.
[4] Timm, N.H. and T.A. Mieczkowski. *General Linear Models: Theory and Applications using SAS® Software*. Cary, NC : SAS Publishing. 1997.
[5] Agresti, A. *An Introduction to Categorical Data Analysis, 2nd edition*. New Jersey, US : John Wiley & Sons. 2007.
[6] SAS Institute Inc. *SAS/STAT® 9.2 User's Guide, Second Edition*. Cary, NC : SAS Institute Inc. 2009.
[7] Collett, D. *Modelling Binary Data. 2nd edition*. London, UK : Chapman & Hall/CRC. 2003.
[8] Saefuddin, A., N.A. Setiabudi and N.A. Achsani. The Effect of Overdispersion on Regression Based Decision with Application to Churn Analysis on Indonesian Mobile Phone Industry. *European Journal of Scientific Research.* 2011. 60(4): 602-610.
[9] Hajarisman, N. and A. Saefuddin. The Beta-Binomial Multivariate Model for Correlated Categorical Data. *Jurnal Statistika*. 2008. 8(1): 61–68.
[10] Kurnia, A., A. Saefuddin and E. Sutisna. Overdispersi dalam Regresi Logistic. *Proc. of Seminar Nasional Statistika, Bogor*. 2002: 11-17.

[11] Hinde, J. and C.G.B. Demétrio. Overdispersion: Models and Estimation. *Computational Statistics and Data Analysis*. 1998. 27: 151-170.
[12] Handayani, D. and A. Kurnia. Mixed Effect Model Approach for Logistic Regression Model with Overdispersion. *Proc. of ICoMS-1*. 2006.
[13] Baggerly, K.A., Deng, L., Morris, J.S. and Aldaz, C.M.. Overdispersed Logistic Regression for SAGE: Modelling Multiple Groups and Covariates. *BMC Bioinformatics*. 2004. 5:144
[14] Lambert, D. and K. Roeder. Overdispersion Diagnostics for Generalized Linear Models. *Journal of the American Statistical Association*. 1995. 90(432): 1225-1236.