M – 11

COMPARISON WILLIAMS METHOD AND BETA-BINOMIAL IN OVERDISPERSION OF LOGISTIC REGRESSION: A CASE OF INDONESIA GENERAL ELECTION DATA 2014

Firman Hidayat, Khairil Anwar Notodiputro, Bagus Sartono
Department of Statistics, FMIPA, Bogor Agricultural University, Indonesia
man_crb14@yahoo.co.id

Abstract

Democratization in Indonesia so far has resulted in increasingly rational voters. The rational voters in each district or city of Indonesia are varied due to many factors. The system of election in Indonesia today is direct election system in which every citizen has freedom to vote the preferred candidates or even not to vote at all. There were 12 political parties participated in the legislative election in 2014, whereas in the presidential election there were two pairs of president and vice-president candidates competed.

This research was aimed to obtain models, at the district level, that properly relate the votes were gained by the two candidates and other variables such as human development index, the results of legislative election as specially coalition of political parties voting results. Since the vote data was binary and showed over-dispersion then a logistics model accounting for over-dispersion was utilized. An over-dispersion problem is present whenever observations which might be expected to correspond to the binomial distribution may have greater variance than $n_i \pi_i (1-\pi_i)$.In this research the William's method and beta-binomial regression were used to overcome the problem. The result showed that the Williams method provided better estimates when was compared to beta-binomial regression.

keyword: Logistic regression, Overdispersion, Williams'Method, Beta-Binomial Regression, General Election

1. INTRODUCTION

1.1 The Background of the Problem

The Systems Election in Indonesia today is the direct election system that provides complete freedom for voters to determine the parties, candidates, and presidential candidate that they will vote.

On the other hand, tightness of the election of membership qualification cause high degree of competition (competitiveness) general election party of 2014 that was followed 12 parties and 2 pairs of candidates for president and vice president. In the president election 2014 there were two Coalitions namely Great Indonesia Coalition (KIH) and the Red and White Coalition (KMP). KIH consists of five parties they are PDIP, Hanura, PKB, Nasdem, and PKPI. While KMP consists of seven parties they are Gerindra, PAN, Golkar, PKS, Demokrat, PPP, and PBB. While the number of pairs of candidates for president and vice president are two couples they are the pair of the first number is Prabowo-Hatta Rajasa and the pair of the second number is Joko Widodo-Jusuf Kalla.

The results of a survey was conducted by the Indonesian Survey Institute in 2011 (LSI 2011) found that the instability of the legislative and presidential elections may be linked to the rationality of voters. The rationality of voters in each district / city in Indonesia vary. This is because each district / city has a different IPM value. The IPM assessment based on some components such as Life Expectancy Rate (AHH), Literacy Rate (AMH), Average the Old

School (RLS), and expenditure per capita that is adjusted (BPS 2008). Human Development Index (IPM) has a role in the election process in Indonesia. Based on the IPM value, the United Nations Development Programme (UNDP) divides human development status into three (3) criteria, namely: Low for IPM of less than 50, Sufficient or Average for IPM value between 50 - 79.9 and High for IPM value of 80 upwards. But there is also the intermediate category subdivide into Lower Intermediate categories (IPM value of 50 to 65.9) and Higher Intermediate (IPM value of 66 to 79.9) (BPS 2008).

In addition because of the rationality of voters, election results can also be determined by the ideological preferences of the voters. This ideological preferences can be reflected by the choice of the party at the time of legislative elections. Therefore, the researcher wants to guess whether voters in Indonesia is a rational voter or not, and whether the voters tend to hold ideological or not with a particular political party coalition to the presidential election 2014 between pairs Prabowo-Hatta Rajasa and Joko Widodo-Jusuf Kalla. By using logistic regression to these problems can be overcome, because the Regression Logistic is regression that is used when the response variable is binary form. However, the logistic regression often occurs anomaly when the variance of the response *y* not follow variety of binom distribution, namely *np (1-p).* If the variance of the response is greater than a variety of binom distribution then this problem in the literature is known as the overdispersion problem. So, in general, overdispersion is a condition in which the diversity of data is greater than the diversity that should be obtained in accordance with the statistical models used. Some causes overdispersion is mistaken in specifying the systematic component, the presence of one or more of data outliers, logistic link function is used to model is not appropriate, the correlation between observations (Kurnia, Handayani 2006). The existence overdispersion cause improper inferences. There are several ways to overcome overdispersion including logistic regression with random effects, Beta-Binomial models, Williams methods, Generalized Estimating Equation (GEE). The method that is used to handle overdispersion in this study is Williams Method and Beta-Binomial. The idea of the method of Williams that match the value of the chi-square person with the expected value approximation. This method gives weighting w_i on observation so that resulting Pearson Chi-Square statistical equations which is approximated by the value of expectation. While Beta-Binomial model is used to accommodate a variety of opportunities response with using Beta distribution (Sutisna 2002).

Therefore it will be obtained how the logistic regression model and handling of overdispersion for the election results and compare the best model between Williams method and beta-binomial method.

1.2 The Aims of the research

The aims of this research are:

1. To make the logistic regression model
2. To handle the overdispersion case with using Williams method and beta-binomial methods
3. To find the best model for the presidential election results

.

2. LITERATURE REVIEWS

2.1 Logistic regression

Logistic regression analysis can be used to determine the effect of several independent variables to categorical response variable

A general model of binary logistic regression is as follows

$$\pi(x_i) = \frac{\exp(\beta_0 + \beta_1 x_1 + ... + \beta_p x_p)}{1 + \exp(\beta_0 + \beta_1 x_i + ... + \beta_p x_p)}$$

(2.1)

The function $\pi(x_i)$ is a nonlinear function that needs to be done logistics transformation to obtain a linear function so that it can be seen the relationship between the dependent variable Y with independent variables X. $\pi(x_i)$ is a link function as g(x) is equation (2.1)is

$$g(x) = \log\left(\frac{\pi(x_i)}{1-\pi(x_i)}\right) \quad (2.2)$$

Then equation (2.1) are substituted in equation (2.2) to obtain:

$$g(x) = \log\left(\frac{\pi(x_i)}{1-\pi(x_i)}\right) = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p \quad (2.3)$$

(Hosmer &Lemeshow, 2000)

2.2. Overdispersion

The Overdispersion occur if actually variance is greater than binomial distribution variance. For the correct model, the value of Pearson's chi-square divided by the degrees of free will be equal to 1. Overdispersion occurs if that value is exceed of 1, and underdispersion occurs if that value is less than 1.

There are two statistics that are used to test of fit model is Pearson's Chi-square and Deviance. This statistics is a function of the residual, i.e the difference between the actual value and predicted values. Pearson residual value for the i-th observation is expressed as follows:

$$\chi^2 = \sum_{i=1}^{N} e_i^2 \text{ dengan } e_i = \frac{\left(y_i - \widehat{\pi}_i n_i\right)}{\sqrt{\widehat{\pi}_i n_i \left(1-\widehat{\pi}_i\right)}}$$

While the residual value deviance for the i-th observation is expressed as follows:

$$d(y_i, \widehat{\pi}_i) = \pm\left\{2\left[y_i \ln\left(\frac{y_i}{n_i \pi_i}\right) + \left(n_i - y_i\right)\ln\left(\frac{(n_i - y_i)}{n_i(1-\pi_i)}\right)\right]\right\}^{1/2}$$

So, deviance can be expressed as follows

$$D = \sum_{i=1}^{n} d(y_i, \widehat{\pi}_i)^2$$

Pearson chi-square and deviance will follow the χ^2 distribution of the degrees of freedom (n-p), and p is the number of parameters in the model are predicted. If the logistic regression models were used for the data is valid, the Pearson Chi-square and deviance will approach the value of the degree of free. This can be explained because the expected value of the χ^2 distribution is equal to the degree of free. If the value of Pearson's Chi-square and deviance greater than the degree of free, then the assumption of binomial variability may not be valid and the data are said to exhibit overdispersion. (Sutisna 2002)

2.3 Williams' Method

Williams method is used to handle overdispersion by giving weight to the parameter estimation of logistic regression, so variance of probability response to be stable when the variance as follows:

$$\text{var}(Y_i) = n_i \pi_i (1-\pi_i)$$

Williams (1982) estimate unknown scale parameter ϕ with the value of Pearson's Chi-square statistics for the full model. Let wi* is the weight of the i-th observation, the Pearson chi-square statistic as follows:

$$\chi^2 = \sum \frac{\omega_i^* (y_i - \widehat{\pi}_i n_i)^2}{\widehat{\pi}_i n_i \left(1 - \widehat{\pi}_i\right)}$$

where $w_i* = 1/\left[1 + \left(n_i - 1\right)\phi\right]$

Expected Value of χ^2 is

$$E\left(\chi^2\right) = \sum_{i=1}^{n} \omega_i^* (1 - \omega_i^* v_i d_i)[1 + \phi(n_1 - 1)]$$

where $v_i = \pi_i n_i (1 - \pi_i)$ and d_i is diagonal element of the variance-covariance matrix of the linear predictor,, $\hat{\eta}_i = \sum \beta_j x_{ji}$. scale parameter ϕ estimated with iteration process. (Saefuddin, Setiabudi 2011). Williams method can only be used for if response variable is in y_i / n_i form (SAS institute 2009).

2.4 Beta-Binomial Regression

Beta-Binomial Regression can be used to handle overdispersion. Beta-Binomial also used to fit the data binomial with modeling probability response variance using the beta distribution. Because the beta distribution has a range (0.1) then it is equal to the probability value. Let π_i distribute as beta distribution with parameter α dan β

π_i ~ Beta (α , β) , α >0 , β >0

Then the probability density function (pdf) is:

$$f(\pi) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \pi^{\alpha - 1} \pi^{\beta - 1} \text{ , where } 0 \le \pi \le 1$$

According Agresti (2002) the distribution of beta-binomial is a combined beta distribution and binomial distribution, i.e Y is assumed to follow the binomial distribution and π follow a beta distribution. Binom beta distribution density function is as follows. The probability density function (pdf) of beta-binomial as:

$$f\left(y; \alpha, \beta\right) = \binom{n}{y} \frac{\Gamma\left(\alpha + \beta\right)}{\Gamma(\alpha)\Gamma\left(\beta\right)} \frac{\Gamma\left(\alpha + y\right)\Gamma\left(n + \beta - y\right)}{\Gamma(\alpha + y + n + \beta - y)}$$

…. (2.4)

The equation (2.4) will have expected value and variance as:

$$E(y \mid n_i) = n_i \left(\frac{\alpha}{\alpha + \beta} \right)$$

$$Var(y \mid n_i) = n_i \frac{\alpha\beta}{(\alpha + \beta)^2} \frac{\alpha + \beta + n_i}{\alpha + \beta + 1}$$

The final model of beta-binomial regression is:

$$Logit(\mu_i) = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip}$$

3. The Methodology of The Research

3.1 Data

This study uses data from the result of 2014 election and the parties vote data result of 2014 according to districts / cities throughout Indonesia from the Commission which derives from www.kpu.go.id site and request directly to the public relations department of KPU on November 14, 2014, As the data from BPS, namely component IPM 2013 (AHH, AMH, RLS, expenditure per capita adjusted) according to district / city from the site www.bps.go.id

Response variable that become concern in this study is binary form that is vote of pair Prabowo -Hatta Rajasa and vote of pair Joko Widodo-Jusuf Kalla, while predictor variables are:

x_1 = Life Expectancy Rate (AHH) with values ranging from 25-85 years

x_2 = Literacy Rate (AMH) with values ranging from 0-100.

x_3 = Average the Old School (RLS) with values ranging from 0-15 years

x_4 = Expenditure per capita that is adjusted (PENG) with values ranging from Rp 360.000 – Rp 732.720

x_5 = Percentage of the political parties votes who are members of KIH (1 -% KMP)

3.2 Analysis Method

Stages of analysis in this study are:

(1) Conduct a test of multicollinearity between independent variables.
To check for multicollinearity can be seen by looking at the value of VIF with the support of Minitab software. If VIF> 10, then there is multicollinearity.

(2) Conduct a logistic regression parameter estimation with using the support of SAS software. The following logistic regression model:

$$g(x) = \log\left(\frac{\pi(x_i)}{1-\pi(x_i)}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5$$

$$g(x) = \log\left(\frac{jokowi - jk}{prabowo - hatta}\right) = \beta_0 + \beta_1 AHH + \beta_2 AMH + \beta_3 RLS + \beta_4 Pengeluaran\_perkapita + \beta_5 KIH$$

(3) Detecting the presence of overdispersion

To detect the presence of overdispersion, it is conducted by dividing the χ^2 *Pearson* statistic value with degrees of freedom. If the ratio χ^2 Pearson / db > 1, then there is a overdispersion. Or if *(p-value> chisq)* $< \alpha = 0.05$ then there is a overdispersion as well.

(4) Conduct modeling of beta-binomial regression and Williams methods

(5) Conduct a test significance of the parameters
To conduct a test significance of the parameters using the Wald test statistic or p-value *(pr > chisq)* $< \alpha = 0.05$ then the parameter is said to be significant

(6) Conduct a model feasibility test
To test whether the model which is produced appropriate or not, therefore it is necessary to test the suitability of the model or the goodness of fit. . The test statistic used is χ^2 *Pearson*.

(7) Selection of the best model
To choose the best model used AIC. The best model is the model that has the smallest AIC value

4. RESEARCH RESULT

In table 1 showed that the predictor variable not occur multi collinear. This is because the value of Variance Inflation Factors (VIF) at predictor variable less than 10, so can be done to analyze of logistic regression.

Table 1. VIF predictor variable

	AHH	AMH	RLS	Pengeluaran Perkapita	KIH
VIF	1.271	2.190	2.569	1.003	1.007

To check overdispersion on logistic regression can be done by sees ratio value Chi square Pearson with its degree of free. In table 2 looked that ratio value *Chi square Pearson* with its degree of free 7375775.29 more greater than 1, so gets to be said that model occurs overdispersion, so have more been handled with William's method and beta-binomial model.

Tabel 2. Criteria overdispersion

Criterion	Value	DF	Value/DF	Pr > ChiSq
Deviance	7610685.58	491	15500.38	<.0001
Pearson	7375775.29	491	15021.95	<.0001

4.1 The Result William's Method

In Table 3, the problem overdispersion have been resolved by the Williams Method. It can be seen from the value of the ratio Pearson Chi-square with degrees free of 1. Or it could be seen from the p-value (Pr> ChiSq) greater than (α =5%) who stated that overdispersion have been resolved.

Tabel 3. The Result William's Method

Criterion	Value	DF	Value/DF	Pr > ChiSq
Deviance	515.3334	491	1.0496	0.2162
Pearson	490.9999	491	1.0000	0.4915

However, in Table 4 show that the estimation of the parameters for expenditure per capita variable (PENG) is not significant, so these variables must be removed from the model. Next will be the repetition of the analysis method of Williams by removing expenditure per capita variable.

Table 4. Estimation of parameters in model 1 Williams Method

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-3.5602	0.7498	22.5471	<.0001
AHH	1	0.0762	0.0109	49.0335	<.0001
AMH	1	-0.0154	0.00363	18.0215	<.0001
RLS	1	-0.0932	0.0270	11.9412	0.0005
PENG	1	0.000029	0.000098	0.0839	0.7721
KIH	1	1.8187	0.2040	79.4933	<.0001

In Table 5 shows that the parameters are already significant for p-value (Pr> ChiSq) is smaller than (α =5%). So that would be the best model is obtained from Williams Method as:

$$\log\left(\frac{jokowi\_jk}{prabowo\_hatta}\right) = -3.5381 + 0.0761AHH - 0.0154AMH - 0.0932RLS + 1.8185KIH$$

Table 5. Estimation parameters of the best model in Williams Method

Analysis of Maximum Likelihood Estimates

Parameter	DF	Standard Estimate	Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-3.5381	0.7452	22.5433	<.0001
AHH	1	0.0761	0.0109	49.0502	<.0001
AMH	1	-0.0154	0.00363	18.0024	<.0001
RLS	1	-0.0932	0.0269	11.9552	0.0005
KIH	1	1.8185	0.2038	79.6195	<.0001

4.2 The Result Beta-Binomial Model

In Table 6 shows that the problem overdispersion has been resolved by beta binomial models. This can be seen from the standard error of each parameter that has a value greater than the standard error of the logistic regression parameters. However, expenditure per capita variable (PENG) is not significant because the p-value (Pr> | z |) is greater than (α =5%). So that these variables must be removed from the model

Table 6. The Result Beta-Binomial Model 1st

Parameter Estimates for 'Beta-Binomial' Model				
Effect	**Estimate**	**Standard Error**	**z Value**	**Pr > |z|**
Intercept	-2.4488	0.7886	-3.11	0.0019
AHH	0.07091	0.01133	6.26	<.0001
AMH	-0.02436	0.003973	-6.13	<.0001
RLS	-0.06773	0.02776	-2.44	0.0147
PENG	0.000020	0.000102	0.20	0.8451
KIH	1.5595	0.2091	7.46	<.0001
Scale Parameter	10.3160	0.6303		

In Table 7 all the parameters of the model has been significant since the beta binomial have p-value (Pr> | z |) is smaller than (α = 5%). So that the model is the best model of beta binomial as:

$$\log\left(\frac{jokowi\_jk}{prabowo\_hatta}\right) = -2.4334 + 0.07083AHH - 0.02434AMH - 0.06771RLS + 1.5592KIH$$

Table 7. The result best models of beta binomial

Parameter Estimates for 'Beta-Binomial' Model				
Effect	**Estimate**	**Standard Error**	**z Value**	**Pr > |z|**
Intercept	-2.4334	0.7847	-3.10	0.0019
AHH	0.07083	0.01132	6.25	<.0001
AMH	-0.02434	0.003971	-6.13	<.0001
RLS	-0.06771	0.02776	-2.44	0.0147
KIH	1.5592	0.2091	7.46	<.0001

Parameter Estimates for 'Beta-Binomial' Model				
Effect	**Estimate**	**Standard Error**	**z Value**	**Pr > |z|**
Scale Parameter	10.3148	0.6302		

4.3 Selection of the Best Model

In Table 8 shows that the best method to solve the overdispersion problem on logistic regression is a method of Williams. This is due to the method of Williams has a smaller AIC value than beta binomial method. So the method of Williams is the best method compared with beta binomial method.

Table 8. The Comparison of AIC Value

Method	AIC
Williams	**8210.747**
Beta binomial	11380.2

The following logistic regression equation obtained from Williams method that is the best in the case of Indonesia in 2014 elections.

$$\log\left(\frac{jokowi\_jk}{prabowo\_hatta}\right) = -3.5381 + 0.0761AHH - 0.0154AMH - 0.0932RLS + 1.8185KIH$$

BIBLIOGRAPHY

Agresti, A., (2002), *Categorical Data Analysis*. New York: John Willey & Sons, Inc.

Badan Pusat Statistik. 2008. *Indeks Pembangunan Manusia 2006-2007*. Jakarta :BPS

Collett. D. 2003. *Modelling Binary Data Second edition*. London, UK : Chapman & Hall/CRC,

Handayani, D. and Kurnia, A. 2006. *Mixed Effect Model Approach for Logistic Regression Model with Overdispersion*, Proc. of ICoMS-1, Bandung.

Hosmer, D.W., Lemeshow, S., 2000. *Applied Logistic Regression*. New York: John Willey and Sons, inc

Irvani, D. 2012. *Pengembangan Interpretasi Model Logit Multinomial dengan Metode Analisis Berbasis Peluang (Studi Kasus: Pemilihan Presiden Tahun 2009)* [Tesis]. Bogor : IPB

Kutner, M.H., Nachtsheim C.J., Neter J.., and Li, W. 2004. *Applied Linier Statistical Models Fifth Edition*. New York: McGraw-Hill, Inc

Lembaga Survey Indonesia. 2011. *Pemilih Mengambang Dan Prospek Perubahan Kekuatan Partai Politik*, Rilis tahun 2011

Saefuddin, A. and Setiabudi, N.A. 2011. *The Effect of Overdispersion on Logistic Regression Analysis of Poverty in Indonesia*, International Journal for Statisticians, Vol. 2(1).

SAS Institute Inc. 2009. *SAS/STAT® 9.2 User's Guide, Second Edition*. Cary, NC : SAS Institute Inc.

Sutisna. E. 2002. *The Problem of Overdispersion in Logistic Regression* [Tesis]. Bogor : IPB

Williams, D.A. 1982. *Extra-Binomial Variation in Logistic Linear Models*. Applied Statistics, Vol. 31 No. 2: 144-148

www.kpu.go.id diakses pada bulan November 2014

www.bps.go.id diakses pada bulan November 2014