Journal of Science and Technology
1 (2), 2007, 245-251

UNBIASED ESTIMATION FOR LOGISTIC REGRESSION

Ecevit Eyduran
eyduran@yyu.edu.tr

Taner Özdemir
taner@yyu.edu.tr

University of Yüzüncü Yıl, Agricultural Faculty, Department of Animal Science, Biometry Genetics Unit, 65080 Van - TURKEY

ABSTRACT

The aim of this study was to determine and reduce the problems that have been encountered in sparse data or separation. For this aim, Firth's Modified Score Procedure, which is more superior to Maximum Likelihood Estimation Method, were performed on a data set regarding psychology. Under some circumstances (sparse data or separation), maximum likelihood estimates of parameters are biased or infinite estimates. Using Firth's Modified Score Procedure or (Firth Type Estimates) may be thought as an originally approach which eliminates or shrinks the first order bias. As a result, it has been suggested in this article that Firth Type Estimates be a reliable method for eliminating separation.

Key words: *Bias shrinking, Firth Type Estimation, Sparse data, Separation.*

ÖZET

Bu çalışmada, psikoloji sparse ya da separation durumunda karşılaşılan problemlerin belirlenmesi ve azaltılması amaçlanmıştır. Bu amaç için, psikolojiyle ilgili bir veri seti üzerine, Maksimum Olabilirlik Tahmin yönteminden daha üstün olan Firth's Modifiye Edilmiş Skor Prosedürü uygulanmıştır. Bazı durumlarda (sparse ya da separation), paramatrelerin Maksimum Olabilirlik tahminleri yanlı ya da sınırsız olmaktadır. Bu yanlılığı gidermek için, orijinal bir yaklaşım olması bakımından Firth's Modifiye Edilmiş Skor Prosedürünün (Firth tipi Tahmin) kullanımı düşünülebilir. Sonuç olarak, bu makalede, separation durumunun elemine edilmesi bakımından Firth tipi Tahmin yönteminin kullanımı önerilmektedir.
Anahtar Kelimeler: *Yansızlık, Firth Tipi Tahmin, Sparse data, Seperation*

1. INTRODUCTION

Logistic regression (LR), a part of generalized linear models, has been widely studied in different scientific fields such as medicine, economy and agriculture recently [1-4]. LR is a multivariate technique where dependent variable is only categorical (generally binary) and independent variables are a mixture of categorical and continuous variables [1-5].

One of the most vital points for investigators is to get reliable estimation to model parameters for statistic methods [6]. When LR is applied for 2x2 (two by two) tables (for instance especially Case-control studies), estimation of parameters obtained from using maximum likelihood is unreliable in problematic data sets where one of the observed frequencies in a 2x2 table is zero [2,3]. Under this condition, estimation of odds ratio is biased or MLE's parameters are infinite estimation when maximum likelihood estimations (MLE's) of slope parameters and their standard errors are too large [3,6]. The phenomenon mentioned above is known generally as Separation [6-9]. MLE's in SAS program for problematic data sets can not performed and the SAS program gives warnings as "**quasi-complete or complete separation**" [3]. On the other hand, converge operations for MLE's in SAS program cannot be carried out [6, 7-9]. When there is separation phenomen in 2x2 contingency tables, as seen in discussion section, the stopping rule for MLE's can be mentioned and the MLE's in this case, the standard errors of parameters are infinite. Besides, [10] reported that bias for MLE's might arise when data were sparse.

Profile penalized log likelihood (PPL) derived from firth's modified score test (FMCT) may be thought as original approach. In the event of bias reduction of MLE's, FMCT suggested by [11] offers researchers finite parameter estimates via penalized maximum likelihood [6, 9]. Besides, in respect of results of simulation studies reported by many researchers Modified Score Estimates are often advantageous to other approaches mentioned above [6, 9]. In addition, as sample size increases, modified score estimates become equal to usual MLE's [6].

Special macros for the advanced method were developed in package programs such as SAS, S-PLUS, and R [7]. It is reported that separation problem and biased estimations were eliminated by the special macros written in 3 package programs [9]. The three macros at different package program developed by [8] were based on profile penalized log likelihood (PPL) (which is also called Firth-type estimates) and Newton-Raphson algorithm was used for PPL parameter estimates in special SAS macro (at the web site: http://www.meduniwien.ac.at/akh/imc/biometrie/programme/fl/). On the other

hand, in sparse sample or in the separation problem mentioned above, it can be implied that Penalized Likelihood Estimation removed $O(n^{-1})$ bias of Maximum Likelihood Estimation [7-13].

However, the applications (use) of three macros founded on FMCT developed by [11] have not been very common. Unfortunately, the applications of estimation methods with PPL including FMCT to Psychology are too rare, but should get some importance. In this context, the aim of present paper was to discuss bias reduction of sparse data obtained by these the area by using only the SAS macro gained to SAS. With the SAS macro, the aim of the study concerning psychology performed on 107 refugees living in Van, Turkey in 2001 was to assess the effect of sex on the occurrence of psychological problem [14].

2. FIRTH'S MODIFIED SCORE PROCEDURE

Maximum likelihood estimates of regression parameters $(\beta_r (r = 1,...,k))$ are found as solutions to the score equations $(\partial \log L / \partial \beta_r \equiv U(\beta_r) = 0)$ where the likelihood function L is. As seen in Eq. 1, [11] suggested basing estimation on modified score equations with the intention of reducing the small sample bias of these estimates.

$$U(\beta_r)* \equiv U(\beta_r) + 1/2 \,\text{trace}\left[I(\beta)^{-1}\{\partial I(\beta)/\partial \beta_r\}\right] = 0 \;(r = 1,...,k) \qquad (1)$$

Where, $I(\beta)^{-1}$ is the inverse of the information matrix evaluated at β. The modified score function $U(\beta)^*$ is associated to the penalized log-likelihood and likelihood functions which can be written as Eq.2 and Eq.3, respectively.

$$\log L(\beta)* \equiv \log L(\beta) + 1/2 \log\left|I(\beta)\right| \qquad (2)$$

$$L(\beta)* = L(\beta)\left|I(\beta)\right|^{1/2} \qquad (3)$$

where, the penalty function, $\left|I(\beta)\right|^{1/2}$, whose influence is not asymptotically meaningful. By using this modification, [11] provides evidence that the $O(n^{-1})$ bias of maximum likelihood estimates $\hat{\beta}$ is removed.

If the modification to a logistic model is applied as follows:

$$\Pr ob(y_i = 1 \mid x_i, \beta) = \pi_i = \left\{1 + \exp\left(-\sum_{r=1}^{k} x_{ir}\beta_r\right)\right\}^{-1} \qquad (4)$$

Then, the score equation may be expressed in Eq.5

$$U(\beta_r) = \sum_{i=1}^{n} (y_i - \pi_i) x_{ir} = 0 \quad (5)$$

Eq.5 is replaced by the modified score equation and rewritten as Eq.6;

$$U(\beta_r)^* = \sum_{i=1}^{n} \{y_i - \pi_i + h_i(1/2 - \pi_i)\} x_{ir} = 0 \qquad (r = 1,\ldots,k) \quad (6)$$

According to Eq.7, Firth-type (FL) estimates $\hat{\beta}$ can be obtained iteratively the usual way until convergence is obtained;

$$\beta^{(s+1)} = \beta^s + I^{-1}(\beta^{(s)}) U(\beta^{(s)})^* \quad (7)$$

where, the superscript (s) refers to the sth iteration [9].

The calculations were performed using a special macro written in SAS statistical package program [15]. The special macro was downloaded from below web site:
http://www.meduniwien.ac.at/akh/imc/biometrie/programme/fl/

3. MOTIVATING EXAMPLE

The aim of this data set was to assess the effect of sex on the occurrence of psychological problem. For this aim, as a contingency table, the data set regarding 107 refugees living in Van, Turkey in 2001 year are presented in Table 1 [14]. Examining Table 1, the number of female and male are 22 and 85, respectively as well as the number of refugees being occurrence and absence of any psychological problem were 76 and 31, respectively. Let's consider that changes in psychology based on sex are (or not) in logit models, that is, suggesting one dependent and explanatory variable such as simple regression. Therefore, we assumed that levels of dependent variable (binary) were Y=1, presence of any psychological problem (PPP) and Y=0, absence of any psychological problem (APP) and levels of explanatory variable are X=1, male and X=0, female. We thought that Female for data set was risk factor of interest (reference category) as independent variable while reference category for dependent variable was presence of psychological problem (PPP).

Table 1. Contingency Table regarding data set.

	Sex (X)	
Psychological Problem (Y)	Female (1)	Male (0)
Presence PPP (1)	22	54
Absence APP (0)	0	31

4. RESULTS AND DISCUSSION

MLE's for the parameters are given in Table 2. As shown in Table 2, it could be said that maximum likelihood Estimations of both parameters and their standard errors in LR were biased and unreliable [3,6,7,8,9,10]. Besides, odds ratio estimate and its 95% Wald Confidence limits for sex were calculated as >999.999 and [<0.001, >999.999], respectively. Looking at Table 2, in the event of being this problem, generally known as "Separation [7,8,9] convergence in logistic procedure of SAS program could not be realized, which were given warnings [3,6,7,8,9] (Figure 1). This could be also identified the stopping rule for MLE's and the standard errors of slopes were found infinite [11,12,13]. Moreover, [10] reported that bias for MLE's might arise when data were sparse. In place of determining the problem, the findings obtained from Table 2 were consistent with those reported by other authors [3,7,8,9,10].

Table 2: MLE's for the parameters on data set I

Parameters	Degrees of Freedom	Estimation of Parameters	Standard Error	Wald Statistics	Probability (P)	Odds Ratio
Intercept (β_0)	1	-13.4563	178.1	0.0057	0.9398	
Female (β_1)	1	12.9013	178.1	0.0052	0.9423	>999.999

WARNING: The maximum likelihood estimate may not exist.
WARNING: The LOGISTIC procedure continues in spite of the above warning. Results shown are based on the last maximum likelihood iteration. Validity of the model fit is questionable.

Figure 1: SAS warnings in separation problem

When Firth's Modified score test as an original approach was applied for data set I, it was reported that profile penalized log likelihood (PPL) Estimation replacing to Maximum Likelihood Estimation. PPL Estimations for data set are presented in Table 3 . Examining in Table 3, FL Estimates and profile-penalized likelihood confidence limits removed $O\left(n^{-1}\right)$ bias of Maximum Likelihood Estimation [7,8,9,11,12,13].

Table 3: FL estimates, profile penalized likelihood confidence limits

Parameters		Estimation of Parameters	Standard Error	Lower 95% CL	Upper 95% CL	P Chi-square
Intercept	1	0.54821	0.22513	0.10696	0.98947	0.0126

(β_0)						
Female (β_1)	1	3.25844	1.47920	1.21894	8.11529	0.0002

NOTE: Confidence interval for Intercept based on Wald method.

FL odds ratio estimates, profile penalized likelihood confidence limits are given in Table 4. As shown in Table 4, Odds ratio value calculated for Firth type Estimation was statistically significant (P<0.01). Besides, the SAS macro removed biased of MLE's given an idea about whether FL odds ratio estimation was significant as being different from Logistic procedure of SAS program and then had not given any warnings in Figure 1. [7,8,9]. Therefore, this is an advantage for researchers who encountered in Separate Problem.

Table 4: FL odds ratio estimates, profile penalized likelihood confidence limits

Effect	Odds Ratio	Profile penalized likelihood confidence limits		
		Lower 95% CL	Upper 95% CL.	Pr > Chi Sq
Female	26.009	3.38361	3345.22	0.0002

5. CONCLUSION

In the event that any one of all cells in contingency table two by two equals was zero, we discussed that many authors described the problem is known as Separation. Encountered in sparse data, it could be said that MLE's for parameters in logistic regression were biased as reported by other authors [7,8,9,10,11,12,13].

It is important for researchers that one of the most fundamental points is to get trustworthy estimation of model parameters for all statistic methods [6].
As a result, it is thought that the SAS macro derived from Firth Modified Score procedure, which is a reliable approach, can be utilized for researchers run into sparse data or separate problem.

REFERENCES

1. Hosmer, D.W., Lemeshow, S. 2000. Applied Logistic Regression (Second Edition). Wiley Series in Probability and Statistics.
2. Sharma, S. (1996). Applied Multivariate Techniques. *John Wiley and Sons, Inc. pages: 317-341.*
3. Allison, P.D. (1999).Logistic Regression Using the SAS System: *Theory and Application. Cary, NC: SAS Institute Inc.*
4. Eyduran, E., Özdemir, T., Çak, B., Alarslan, E. 2005. Using of logistic regression in Animal Science. *Ansinet, Journal of Applied Sciences 5(10): 1753-1756.*
5. Agresti, A. (2002). Categorical Data Analysis. *John Wiley and Sons, Inc.*

6. Bull, S.B., Mark, C., Greenwood, C.M.T. (2002). A modified score function estimator for multinomial logistic regression in small samples. *Computational Statistics and Data Analysis, 39: 57-74.*
7. Heinze G, Schemper M (2002). A solution to the problem of separation in logistic regression. *Statistics in Medicine 21: 2409-2419.*
8. Heinze G, Ploner M (2003). Fixing the nonconvergence bug in logistic regression with SPLUS and SAS. *Computer Methods and Programs in Biomedicine* 71: 181-187.
9. Heinze G, Ploner M (2004). Technical Report 2/2004: A SAS-macro, S-PLUS library and R package to perform logistic regression without convergence problems. Section of Clinical Biometrics, Department of Medical Computer Sciences, Medical University of Vienna, Vienna, Austria. http://www.meduniwien.ac.at/akh/imc/biometrie/programme/fl/
10. Greenland, S. (2000). Small-sample bias and corrections for conditional maximum likelihood odds ratio estimators. *Biostatistics 1,1, pages: 113-122.*
11. Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika* 80, 27–38.
12. Firth, D. (1992a). Bias reduction, the Jeffreys prior and GLIM. In Fahrmeir, L., Francis, B., Gilchrist, R., & Tutz, G., editors, *Advances in GLIM and Statistical Modelling, pages 91–100. Springer-Verlag, New York.*
13. Firth, D. (1992b). Generalized linear models and Jeffreys priors: an iterative weighted least-squares approach. In Dodge, Y. & Whittaker, J., editors, *Computational Statistics, volume 1, pages 553–557, Heidelberg. Physica-Verlag.*
14. Özdemir, T. (2001). The problems of asylum seeker waiting for refugees or accepted as refugees by Union National High Commissary Refugees in Van, Turkey. University of Yüzüncü Yıl, Institution of Natural and Applied Sciences, *Master Thesi, pages:1-16.Van-Turkey*
15. SAS, (2005). SAS Institute Inc.Version 8, Cary, NC, USA.