# Inference for the coefficient parameters of the logistic regression from a small sample

Kamakura, Toshinari
*Chuo University, Department of Industrial & Systems Engineering*
*1-13-27 Kasuga, Bunkyo-ku*
*Tokyo 112-8551,Japan*
*E-mail: kamakura@indsys.chuo-u.ac.jp*

## 1 Introduction

The logistic model is widely used in the field of medical and industrial applications for evaluating the risk factors adjusting confounding factors based on the event occurrence data and covariates. More than two thousand articles were published in 1999 according to Ryan (2000), but as much less than 1% of the large number of papers on logistic regression appear in statistics journals, despite the fact that there are many unsolved research issues.

In this paper we investigate the Wald test of the regression parameter in case of a small sample. Hauck and Donner (1977) pointed out that behavior of the Wald and the likelihood ratio test do not coincide. Furthermore, in case of the probability of occurrences of the event is so high and low, it is known the phenomenon of separation or monotone likelihood is observed in the fitting process of a logistic model and the Wald test sometimes gives rise to very conservative results. In this case the likelihood test is also conservative in our simulation studies. Separation primarily occurs in small samples with several unbalanced and highly predictive risk factors (Albert and Anderson, 1984; Heinze and Schemper, 2002). We investigate separation or quasi-separation for dataset and calculate probability of separation or quasi-separation as the function of event occurrence probability and test and study the property of p-values of the tests.

We proposed new method based on the bootstrap resampling techniques and compare the true p-values for the likelihood ratio test, Wald test, and other testing methods (Ohkura and Kamakura, 2011). Simulations studies illuminates that our new method keeps nominal p-values even for the high or low event probabilities. The proposed method is based on the bootstrap technique and the modification of the loglikelilhood by Firth (1993). The Firth (1993) method has a good property that it gives the bias reduction of the maximum likelihood estimates and the stable estimates are obtained even for the nearly quasi-separation case. Combination of the bootstrap method and the Firth method will give a good performance even for moderately high probability of event occurrences.

## 2 Problems of testing regression parameters

Firstly, we consider the case that a single regression variable is included in the logistic regression models. We assume that the binary observation set is $\{y_1, \ldots, y_n\}$ andthat its corresponding covariate set is $\{x_1, \ldots, x_n\}$. We are interested in testing the slope parameter $\beta$ included in the regression model:

$$p_i = P\{Y_i = 1\} = \frac{\exp(\alpha + \beta x_i)}{1 + \exp(\alpha + \beta x_i)} \quad (i = 1, \ldots, n). \tag{1}$$

The parameter $\beta$ and $\alpha$ are sometimes called the structural parameter and the nuisance parameter individually. In this case the statistical hypotheses are $H_0 : \beta = 0$ and $H_A : \beta \neq 0$. In the field of application to the risk factors for biomedical or engineering data set the Wald test is frequently used

based on the following log-likelihood.

$$\log L = \log \left[ \prod_{i=1}^{n} \left\{ p_i^{y_i} (1-p_i)^{1-y_i} \right\} \right]. \tag{2}$$

The ML estimates are basically calculated by some iterative algorithms like a Newton method. The log-likelihood has a good property that it is concave and we can easily obtain ML estimates except in case of a small sample. The Fisher information matrix is as follows:

$$I = \begin{pmatrix} \sum_{i=1}^{n} (1-p_i)\, p_i & \sum_{i=1}^{n} (1-p_i)\, p_i x_i \\ \sum_{i=1}^{n} (1-p_i)\, p_i x_i & \sum_{i=1}^{n} (1-p_i)\, p_i x_i^2 \end{pmatrix}. \tag{3}$$

Then the asymptotic variance of the ML estimator $\hat{\beta}$ is calculated as

$$Avar(\hat{\beta}) = \frac{\sum_{i=1}^{n} (1-p_i)\, p_i}{(\sum_{i=1}^{n} (1-p_i)\, p_i) \sum_{i=1}^{n} (1-p_i)\, p_i x_i^2 - (\sum_{i=1}^{n} (1-p_i)\, p_i x_i)^2}. \tag{4}$$

As ML estimate cannot be expressed in explicit form, the Wald statistic is neither obtained in a closed form. Under null hypothesis $H_0 : \beta = 0$ the equation (4) becomes

$$\mathrm{Avar}(\hat{\beta})\Big|_{\beta=0} = \frac{1}{p(1-p)} \times \frac{1}{n \times \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2}. \tag{5}$$

The equation (5) is interpreted to give the larger asymptotic variance when the baseline occurrence probability is small or large and the sample variance of the covariate is small. Our simulation result supports that the Wald test give rise to heavy conservativeness for significance test of the null hypothesis. Pooi (2003) and Hauck and Donner (1977) also indicate that Wald test may result in conservative testing for a small sample. Figure 1 is the simulation result that comes from Okura and Kamakura (2011). Considering symmetry for occurrence probability around $p = 0.5$ the upper part result is shown. We can observe very small rejection probability for the larger occurrence probability.

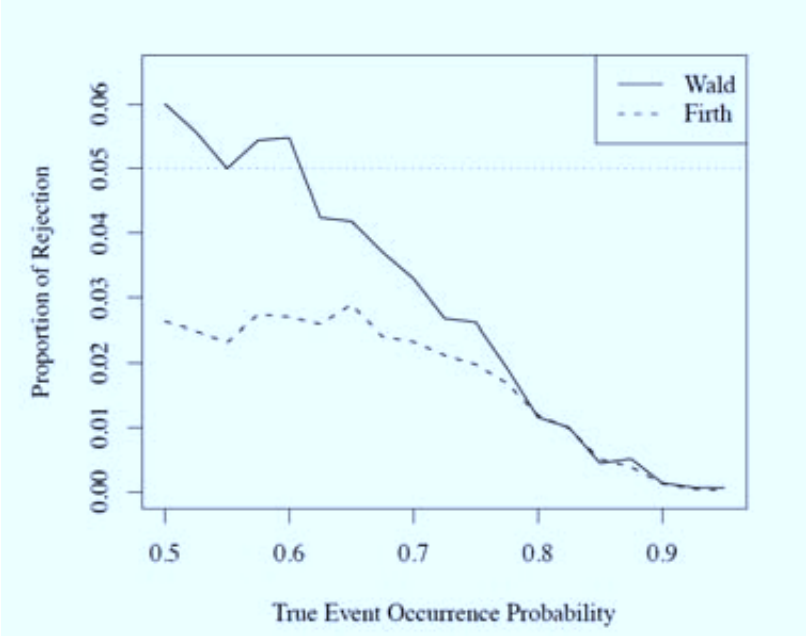


Figure 1: Conservativeness of the Wald test from Okura and Kamakura (2011)

## 3 Convergence problem

Though the log-likelihood of the logit model have good property that its Hessian matrix is concave, we cannot sometimes obtain the ML estimates especially for small samples. This is because that the ML estimates has the infinity solution or numerical solutions do not converge. The commercial statistical package shows us some messages of no convergence results. However, R language give us no messages of convergence.

The case-control study of Foxman et al. (1997) examines urinary tract infection in related to age and contraceptive use. The data set consists of 130 college women with urinary tract infections and 109 uninfected controls. The data set include the binary covariates age (age), oral contraceptive use (oc), condom use (vic), lubricated condom use (vicl), spermicide use (vis) and diaphragm used (dia) The following list is an example of no convergence. As for the covariate dia the regression coefficient

```
> ans=glm(case~.,family=binomial(link=logit),data=sex2)
> summary(ans)

Call:
glm(formula = case ~ ., family = binomial(link = logit), data = sex2)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2850  -1.0811   0.3907   1.1542   1.6701

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.12835    0.49095   0.261  0.79377
age          -1.16439    0.43452  -2.680  0.00737 **
oc           -0.07356    0.44967  -0.164  0.87005
vic           2.40593    0.56953   4.224 2.40e-05 ***
vicl         -2.24620    0.56433  -3.980 6.88e-05 ***
vis          -0.82011    0.42158  -1.945  0.05174 .
dia          16.73423  799.42507   0.021  0.98330
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

estimate seems not to be converged. The estimate of the standard error is very large and this indicates the no convergence. Very large standard error considered to be good indicator for no convergence.

## 4 Bootstrap choice

The bootstrap method is a kind of useful technique especially for the case of difficulties in deriving null distribution. However, if we use the bootstrap method for small sample case, separable sample frequently occurs. For the case of separable or quasi-separable sample the ML estimates may diverge, and so we cannot obtain the ML estimates. This is a problem of application of the bootstrap method to the small samples for logistic regression model. In our experience (Okura and Kamakura, 2011) the Firth method (Firth, 1993) is very efficient if we jointly use the bootstrap method. This was devised by Firth (1993) for bias reduction of the regression parameter; the log likelihood is adjusted as the

following.

$$\log L^* = \log L + \frac{1}{2}|I|. \tag{6}$$

On calculation of bootstrap iteration we considered the following three choices.

- The case of non-convergence is deleted.(ML)
- The case of non-convergence is replace by the value one.(ML)
- The Firth method is applied for calculations because of its good property of convergence. (Heinze and Schemper, 2002)

Even for the high probability occurrences the simulation result shows us high performance of the method of combination of adjusted log-likelihood and bootstrap method.

## 5 Discussion

In case of the probability of occurrences of the event is so high and low, the phenomenon of separation or monotone likelihood is observed in the fitting process of a logistic model and the Wald test sometimes give rise to very conservative results. In this article we investigated separation or quasi-separation for dataset and calculate probability of separation or quasi-separation as the function of event occurrence probability. We proposed new method based on the bootstrap resampling techniques and compared the true p-values for Wald test, Firth test and other testing methods. Simulations studies illuminate that our new method keeps nominal p-values even for the high or low event probabilities in small samples.

### REFERENCES (RÉFERENCES)

[1] Albert, A. and Anderson. J. A. (1984). On the existence of maximum likelihood estimates in logistic regression models, Biometricka, 71, 1-10.
[2] Firth, D. (1993). Bias reduction of maximum likelihood estimates, Biometrika, 80, 27-38.
[3] Foxman, B., Marsh, J., Gillespie, B., Rubin,N, K.., Koopman,J. S., & Spear, S. (1997). Condom use and first-time urinary tract infection. Epidemiology 8, 637-641.
[4] Hauck, Jr., W. W. and Donner, A. (1977). Wald's test as applied to hypotheses in logit analysis, JASA, 72, 851-853.
[5] Heinze, G.. and Schemper, M. (2002). A solution to problem of separation in logistic regression, Statist. Med.,21, 2409-2419.
[6] Ohkura, M. and Kamakura, T. (2011). Test for a regression parameter in a logistic regression model under the small sample size and the high event occurrence probability, Japanese Applied Statistics (in Japanese), 40, 41-51.
[7] Pooi, H. A. (2003). Performance of the likelihood Ratio Test When fitting Logistic Regression Models with Small Samples, Communications in Statistics: Simulation and Computation, 32, 411-418.
[8] Ryan, T. P. (2000). Some issues in logistic regression, 29, 2019-2032.