http://www.pjbs.org

PJBS

ISSN 1028-8880

Pakistan Journal of Biological Sciences

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Pakistan Journal of Biological Sciences 4 (9): 1180-1183, 2001

Bayesian Inference of Binary Logistic Regression Model for Assessing Erythrocyte Sedimentation Rate

[1]Mehmet Ali Cengiz, [2]Yüksel Bek and [1]Rezan Yilmaz
[1]Department of Statistics, University of Ondokuz Mayis, Samsun, Turkey
[2]Department of Biostatistics, University of Ondokuz Mayis, Samsun, Turkey

Abstract: Response variables having two possible categories are called binary variables. We often describe two possible categories as the terms of disease and healthy. Binary response data are modelled using the binomial distribution while binary data may be assumed to have the Bernoulli distribution which is a special case of the binomial distribution. This paper investigates logistic regression model to improve the accuracy of predictions and decisions, in the specific context of assessing erythrocyte sedimentation rate. The analysis is enhanced further by adopting a Bayesian approach.

Key words: ESO, Logistic model, Bayesian inference, Binomial distribution

Introduction

Erythrocyte Sedimentation Rate (ESR) is a non-specific marker of illness. ESR is the rate at which red blood cells settle out of suspension in blood plasma, when measured under standard conditions. The ESR is a non-specific test and so is difficult to interpret. Recent trials of ESR have demonstrated no value in screening asymptomatic individuals, because not only is the number of abnormals low but also in most cases the abnormal test returns to normal over several months without any significant diagnosis being made (Sox and Liang, 1986). ESR increases if the levels of certain proteins in the blood plasma rise, such as in rheumatic diseases, chronic infections and malignant diseases. This makes the determination of ESR one of the most commonly used screening tests performed on samples of blood.

Such a research was made by the Institute of Medical Research, Kuala Lumpur, Malaysia. They used ESR related to two plasma proteins, fibrinogen and γ-globulin, both measured in gm/l, for a sample of thirty-two individuals. The original data were presented by Collett and Jemain (1985) and were reproduced by Collett (1996), who classified the ESR as binary (0 or 1). Since the ESR for a healthy individual should be less than 20 mm/h and the absolute value of ESR is relatively unimportant, a response of zero signifies a healthy individual ($ESR < 20$) while a response of unity refers to an unhealthy individual ($ESR \geq 20$). Their aim of a statistical analysis of these data was to determine the strength of any relationship between the probability of an ESR reading greater than 20 mm/h and levels of two plasma proteins. In modelling the dependence of the disease state on these two explanatory variables, they use a number of possible linear logistic models with Classical inference.

The aim of present study is to determine the probability of an ESR using Bayesian inference, in place of Classical inference, and to compare these two approaches and then to present new approaches in assessing the probability of ESR. We use logistic regression and the discriminant analysis. Logistic regression is commonly used when the independent variables include both numerical and nominal measures and the outcome variable is binary, or dichotomous, having only two values. It requires no assumptions about the distribution of the independent variables. Another advantage is that the regression coefficient can be interpreted in terms of relative risks in cohort studies or odds ratios in case-control studies. The discriminant analysis is similar to logistic regression in that it is used to predict a nominal or categorical outcomes.

Discriminant analysis differs from logistic regression in that the independent variables follow a multivariate normal distribution, so it must be used with caution if some independent variables are nominal. In many research situations, either logistic regression or discriminant analysis can be used, depending on how the problem is defined.

The last two decades have seen increased usage of another principal inference to statistical analysis. This is Bayesian inference, based on the famous published posthumously by the Reverend Thomas Bayes in 1763 and reproduced in Press (1989). In Bayesian inference the numerical values allotted to probabilities do not relate to long-run frequencies and attempt is made to account for prior knowledge by quantitative measurement. Bayesian inference conditions on the data and integrates over the parameters to evaluate the probabilities rather than estimates, hypothesis tests and confidence. The process of inference requires the evolution of further integrals and the selection of appropriate prior.

In this paper we present suitable prior distributions. In some practical applications there is very little prior information available. In these circumstances a vague prior can be used. The standard choice over recent years has been the invariant prior proposed by Jeffreys (1939). The other suitable priors may be Uniform and Improper, which are described by Bernardo and Smith (1994). The evaluation of integrals may be difficult analytically but numerical methods can be used to overcome this difficulty. Dunsmore (1976) considered an asymptotic Bayesian approach to prediction analysis. We use this approach and modify to binary data.

The linear logistic model to binomial data: Suppose that we have n binomial observations of the form , y_i, $i = 1,\ldots,n$ where $E(y_i) = p_i$ and p_i is the success probability corresponding to the ith observation. The linear logistic model for the dependence p_i on the values of the k explanatory variables x_{1i}, x_{2i},.....x_{ki}, associated with that observation, is

$$\text{logit}(p_i) = \log(p_i/(1-p_i)) = \beta_0 + \beta_{1x1i} + \ldots + \beta_k x_{ki} \quad (1)$$

After some rearrangement,

$$p_i = \frac{1}{1 + \exp(-\beta_0 - \beta_1 x_{1i} - \ldots - \beta_k x_{ki})} \quad (2)$$

If we write

$$\theta_i = \beta_0 + \beta_1 x1_i + \beta_2 x_{2i} + \ldots + \beta_k x_{ki}, \text{ then ,}$$

$$p_i = \frac{1}{1 + e^{-\theta_i}} \quad (3)$$

In order to fit a linear logistic model to a given set of data, unknown parameters must first to be estimated. Several approaches may be used to estimate the unknown parameters.

In classical approach, these parameters are estimated using the methods of maximum likelihood. The likelihood function is given by

$$L(\beta_i;y_i)=\prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{1-y_i} \qquad (4)$$

The problem is to obtain estimations of parameters which maximise

$$l(\beta_i;y_i)=\sum_{i=1}^{n} y_i \log(p_i) + (1-y_i)\log(1-p_i)$$

We do not work on obtaining Maximum likelihood estimations in this paper. We use Bayesian inference to obtain parameter estimations.

Bayesian inference: Suppose we have some training data, $D=\{(Z_i,y_i);i=1, \dots, n\}$ which consist of observed response vectors y_i and matrices of explanatory variables Z_i. Typically, we will observe Z_{n+1} for a new individual, and our aim is to predict the response vector y_{n+1}. The conditional distribution of y_i given Z_i is assumed known as a function of unknown parameters contained in a vector β. The posterior predictive distribution of y_{n+1} given Z_{n+1} and the data D is given by

$$f(y_{n+1}|Z_{n+1}, D)=\int_{\beta} f(y_{n+1}|Z_{n+1}, \beta)\, f(\beta|D)d\beta \qquad (5)$$

This is the basic equation of Bayesian predictive inference in regression analysis. We may re-express equation (3) as

$$f(y_{n+1}|Z_{n+1},D) = \frac{\int_{\beta} f(y_{n+1}|Z_{n+1},\beta)\, L(\beta;D)\, f(\beta)\, d\beta}{\int_{\beta} L(\beta; D)\, f(\beta)\, d\beta} \qquad (6)$$

or

$$f(y_{n+1}|Z_{n+1},D) \;\alpha\; \int_{\beta} f(y_{n+1}|Z_{n+1},\beta)\, L(\beta;D)\, f(\beta)\, d\beta$$

where $L(\beta;D)$ is the likelihood function and $f(\beta)$ is the prior distribution. The crucial task is to evaluate at least the numerator integral in equation (6), as a function of y_{n+1} or numerically for different values of y_{n+1}. Generally, the required integrations are not feasible analytically and approximation methods are needed.

Dunsmore (1976) considered an asymptotic Bayesian approach to prediction analysis. If we expand $f(y_{n+1}|Z_{n+1},\beta)$ in equation (6) about the maximum likelihood estimate β of by Taylor`s theorem, we obtain

$$f(y_{n+1}|Z_{n+1},D)=$$

$$\int_{\beta}\{f(y_{n+1}|Z_{n+1},\hat{\beta}) + (\beta-\hat{\beta})f'(y_{n+1}|Z_{n+1},\hat{\beta})$$

$$+\frac{1}{2}(\beta-\hat{\beta})^2 f''(y_{n+1}|Z_{n+1},\hat{\beta})\}f(\beta|D)d\beta$$

$$\approx f(y_{n+1}|Z_{n+1},\beta)+\frac{1}{2} I_{\beta}^{-} f''(y_{n+1}|Z_{n+1},\beta) + \qquad (7)$$

where I_{β} is the unexpected analogue of Fisher`s measure of information,

$$I_{\beta}=-l''(\beta;D)$$

A first order approximation and second order approximation to the predictive distribution are then obtained by truncating the series in equation (7). In this paper we obtain a first order approximation and use it to obtain posterior predictive probabilities for ESR data.

Analysis and Results

As we mentioned earlier, the ESR for a healthy individual should be less than 20 mm/h and since the absolute value of ESR is relatively unimportant. A response of zero signifies a healthy individual (ESR < 20), while a response of unity refers to an unhealthy individual (ESR ≥ 20). Explanatory variables are the amounts of fibrinogen and γ-globulin. Data are given in Table 1.

The linear logistic model for the dependence on the values of the two explanatory variables x_{fi} and x_{gi};

$$\text{logit}(p_i)=\log\left(p_i/(1-p_i)\right)=\beta_0+\beta_1 x_{fi}+\beta_2 x_{gi}$$

(i = 1,2,...,32).

Where
x_{fi}: the amount of fibrinogen for individual i,
x_{gi} : the amount of γ-globulin for individual i,
y_i : the response variable (ESR).

After some rearrangement,

$$p_i=\frac{1}{1+\exp\left(-\beta_0-\beta_1 x_{fi}-\beta_2 x_{gi}\right)}$$

Probability function is

$$f(y_i)=P_i^{y_i}\,(1-p_i)^{1-y_i},\ (i=1,2,\dots,32).$$

Maximum likelihood function is

$$l(\beta_i;y_i)=\sum_{i=1}^{32} y_i \log(p_i)+(1-y_i)\log(1-p_i).$$

Suitable priors are given in Table 2.

Using Improper prior given in Table 2, we have posterior function as follows:

$$f(B_i;y_i)=$$

$$\left[\sum_{i=1}^{32} y_i\log(p_i)+(1-y_i)\log(1-p_i)\right]$$

$$x\left[p_i^{-1}\,(1-p_i)^{-1}\right]$$

Table 1: The leves of two plasma proteins and the value of a binary response variable.

n	y_i	x_{fi}	x_{gi}	n	y_i	x_{fi}	x_{gi}
1	0	2,52	38	17	1	3,53	46
2	0	2,56	31	18	0	2,68	34
3	0	2,19	33	19	0	2,60	38
4	0	2,18	31	20	0	2,23	37
5	0	3,41	37	21	0	2,88	30
6	0	2,46	36	22	0	2,65	46
7	0	3,22	38	23	1	2,09	44
8	0	2,21	37	24	0	2,28	36
9	0	3,15	39	25	0	2,67	39
10	0	2,60	41	26	0	2,29	31
11	0	2,29	36	27	0	2,15	31
12	0	2,35	29	28	0	2,54	28
13	1	5,06	37	29	1	3,93	32
14	1	3,34	32	30	0	3,34	30
15	1	2,38	37	31	0	2,99	36
16	1	3,15	36	32	0	3,32	35

Table 2: Prior distributions

Distributions	Priors	Priors for Bernoulli distribution
Uniform	$f(\beta(a,b) = \begin{cases} 1/b\text{-}a; & \beta \in (a,b) \\ 0; & \beta \notin (a,b) \end{cases}$	1 to (a,b) = (1,1)
Jeffrey's	$f(\beta) = \sqrt{-E\left\{\frac{d^2 \log f(y\|\beta)}{d\beta^2}\right\}}$	$\left[pi^{-1/2}(1\text{-}pi)^{1/2}\right]$
Improper	$f(\beta) = -E\left\{\frac{d^2 \log f(y\|\beta)}{d\beta^2}\right\}$	$\left[pi^{-1}\,1\text{-}pi)^{-1}\right]$

We apply five cases to our data:

Case 1: Discriminate analysis
Case 2: Linear Logistic regression with Classical inference (using Maximum Likelihood Method)
Case 3: Linear Logistic Regression with Bayesian inference (using Uniform prior and First Order approximation)
Case 4: Linear Logistic Regression with Bayesian inference (using Jeffreys's prior and First Order approximation)
Case 5: Linear Logistic Regression with Bayesian inference (using Improper prior and First Order approximation)

For our model with different inferences and priors, we apply the method using FORTRAN computer programs and subroutines from the NAG library to obtain approximate posterior predictive distributions as given by equation (8). The estimations of parameters are

$$\hat{\beta} = -10.710, \qquad \hat{\beta} = 1.620 \text{ and } \hat{\beta} = 0137.$$

The linear logistic model is

$$\text{logit}(p_i) = \log(p_i/(1\text{-}p_i))$$
$$= -10.710 + 1.620x_{fi} + 0.137x_{gi}$$

Approximate posterior predictive probabilities for the Case 5 are given in Table 3. "*" denotes misclassified ESR. ESR's for three individuals (14,15 and 23) are misclassified.

Table 3: Posterior predictive probabilities for case 5.

n	ESO posterior prediction	n	ESO posterior prediction	
1	0,19	17	0,79	
2	0,09	18	0,15	
3	0,07	19	0,21	
4	0,05	20	0,12	
5	0,47	21	0,13	
6	0,14	22	0,47	
7	0,43	23	0,21	*
8	0,11	24	0,11	
9	0,43	25	0,26	
10	0,29	26	0,06	
11	0,11	27	0,05	
12	0,05	28	0,06	
13	0,93	29	0,51	
14	0,28*	30	0,23	
15	0,14 *	31	0,28	
16	0,34	32	0,37	

We use a criteria to assess our predictive accuracy for each model. This is a binary loss function, corresponding to the

percentage of correct classifications based on cross validation of the training data set with a default classification threshold of 0.5. Table 4 gives the percentages of correct classifications for all cases.

Table 4: Predictive accuracy results for Cases 1-5

Cases	The percentages of correct classification	Number of individuals missclassified
1	75	5,7,9,15,16,22,23,32
2	87.5	14,15,23,29
3	87.5	14,15,23,29
4	87.5	14,15,23,29
5	90.625	14.15,23

The coefficients of x_{fi} (fibronogen) and x_{gi} (γ-globulin) ($i=1,\ldots,32$) in (9) can be interpreted as log odds ratio. Consider the ratio of the odds of disease for an individual for whom the value $x_{fi}+1$ is recorded relative to one for whom the value x_{fi} is obtained. This is given by

$$\Psi_{fi} = \frac{\exp\left(\beta_0 + \beta_1\left(x_{f_i} + 1\right)\right)}{\exp\left(\beta_0 + \beta_1 x_{f_i}\right)} \quad = \quad \mathbf{exp\ (\beta_1)}$$

and so $\hat{\beta}_1$ is the estimated change in the logarithm of the odds ratio when (fibrinogen) increased by one unit. Similarly,

$$\Psi_{g_i} = \exp\left(\beta_2\right)$$

and so $\hat{\beta}_2$ is the estimated change in the logarithm of the odds ratio when (γ-globulin) increased by one unit. When x_{fi} and x_{gi} are increased by *r* units, the estimated changes in the odds are given in Table 5.

Table 5: The estimated changes in the logarithm of the odds ratio

r	Ψ_f	Ψ_g
1	5.05	1.15
2	25.53	1.31
3	129.02	1.51
4	651.97	1.73
5	3294.47	1.98

Conclusions: This paper described and discussed the properties, and an application, of binary logistic regression model, suggesting simplifications and suitable approximations. It considers modifications to the basic model, through using different inferences and prior distributions. The results clearly demonstrate that these binary logistics model facilitates reasonably good assessments of ESO.

It should be emphasized that Logistic models give better percentages of correct classifications than discriminate analysis. Logistic model with Classical inference gives the same percentage as Bayesian inference with Uniform and Jeffreys priors. We have about 91% of correct classifications for Bayesian inference with Improper prior. This percentage is much higher than others. This is a very good success rate. We hope to improve the results further, by considering other suitable priors, covariates and approximation techniques.

We obtained the estimated changes in log odds ratio. Regarding the odds ratio as an approximation to the relative risk, this can be interpreted as the approximate change in the risk of the disease for every increase in the amounts of fibrinogen and γ-globulin.

References

Bernardo, J. M. and A. F. M. Smith, 1994. Bayesian theory. John Wiley & Sons.

Cengiz, M.A., 1997. Bivariate Logistic Regression Analysis. Technical Report, University of Salford, MCS-9711.

Collett, D., 1996. Modelling Binary Data. Chapman & Hall.

Collett, D. and A. A. Jemain, 1985. Residuals, outliers and influential observations in regression analysis. Sains Malaysiana, 14: 493-511.

Copas, J. B., 1988. Binary Regression models for contaminated data. J. Royal Stat. Soc., B50: 225-265.

Cox D. R. and E. J. Snell, 1989. Analysis of Binary Data. Chapman & Hall.

Crowder M. and T. Sweeting, 1989. Bayesian Inference for a bivariate binomial. Biometrika, 76: 599-604.

Dunsmore, I. R., 1976. Asymptotic prediction analysis. Biometrika, 63, 3: 627-630.

Percy, D. F., 1992. Blocked arteries and multivariate regression. Biometrics, 48: 683-693.

Press, S. J., 1989. Bayesian Statistics: principles, models and applications. New York: Wiley.

Sox, H.C. and M. H. Liang, 1986. The erythrocyte sedimentation rate. Ann. Int. Med., 515-523.