LOGISTIC RIDGE REGRESSION FOR CLINICAL DATA ANALYSIS (A CASE STUDY)

E. VÁGÓ [1]-S. KEMÉNY[1]*

[1]*Department of Chemical Engineering, Budapest University of Technology and Economic H-1111, Budapest, Műegyetem rakpart 3, Hungary*
(phone: +36-1-463-2209; fax: +36-1-463-197)

**e-mail: kemeny@mail.bme.hu*

(Received 10th Sep 2005, accepted 10th Oct 2006)

Abstract. This paper focuses on regression with binomial response data. In these cases logit regression is the most used model. An example is a retrospective biomedical problem, where multicollinearity occurs, thus the variances of the estimated parameters are large.
In this paper we propose to apply the ridge method to the maximum likelihood estimation of the logit model parameters.
The efficiency of the proposed technique was investigated using a biomedical data set. A random sampling technique was used to study the effect of sample size on the ML and the logistic ML estimation.
Keywords: *logit, multicollinearity, bootstrap, restless legs*

Introduction

Logit regression is a widely used method for categorical response data. A typical area of application is biomedical studies, but there are other areas like the prediction of loan returning behaviour of bank clients. A good example is the investigation of the occurrence of a disease (yes/no) as related to different characteristics of the patients.

With logit regression the binary response (y_i) at the i-th setting of independent (regressor) variables is considered as a binomial (Bernoulli) random variable with p_i parameter.

$$y_i \sim Binomial(p_i) \qquad \text{(Eq. 1.)}$$

The *logit* is the link function to the linear predictor [1].

$$\text{logit}[p_i(\boldsymbol{x}_i)] = \log\frac{p_i(\boldsymbol{x}_i)}{1-p_i(\boldsymbol{x}_i)} = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_h x_{ih} \ , \qquad \text{(Eq. 2.)}$$

where x_{ij} is the value of the j-th independent variable (j=1…h) at the i-th measurement point, and β_j is the coefficient of the j-th independent variable. From Eq. 2. the following model relates the probability of occurrence with the regressor variables

$$p_i(\boldsymbol{x}_i) = \frac{\exp(\boldsymbol{x}_i\boldsymbol{\beta})}{1+\exp(\boldsymbol{x}_i\boldsymbol{\beta})} \qquad \text{(Eq. 3.)}$$

where

http://www.ecology.uni-corvinus.hu ● ISSN 1589 1623

$$\boldsymbol{x}_i\boldsymbol{\beta} = \beta_0 + \beta_1 x_{i1} + ... + \beta_h x_{ih} \qquad \text{(Eq. 4.)}$$

p_i is the probability of one of the two specific outcomes at the i-th setting of independent (regressor) variables.

When the number of observations at each $\boldsymbol{x}_i$ is not small weighted least squares estimation method can be used [2]. In case of small sample sizes or ungrouped data (n_i=1 for each i) maximum likelihood estimation is applied. This paper focuses only on the latter case.

Maximum likelihood estimation to logit model

Maximum likelihood estimators are obtained by maximizing the logarithm of the likelihood function [1]:

$$\log(L(\boldsymbol{X},\boldsymbol{\beta})) = l(\boldsymbol{X},\boldsymbol{\beta}) = \sum_{i=1}^{n} y_i \log(p_i) + \sum_{i=1}^{n} (1-y_i)\log(1-p_i) \rightarrow \max(\boldsymbol{\beta}) \qquad \text{(Eq. 5.)}$$

where n is the number of observations and $\boldsymbol{X}$ is an $n \times m$ matrix of the independent variable. The estimator is asymptotically unbiased.

Differentiating Eq. 5. with respect to $\boldsymbol{\beta}$ we obtain [3]:

$$\frac{\partial l(\boldsymbol{X},\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \boldsymbol{X}^T(\boldsymbol{Y} - \boldsymbol{p}) \qquad \text{(Eq. 6.)}$$

where $\boldsymbol{Y}$ is an $n \times 1$ vector of observable dependent variables. The maximum likelihood estimator of $\boldsymbol{\beta}$ is obtained by setting the right hand side of these equations equal to zero and then solving them simultaneously and iteratively. Since $\hat{\mathbf{Y}} = \hat{\mathbf{p}}$, Eq. 6. will satisfy

$$\boldsymbol{X}^T(\boldsymbol{Y} - \hat{\boldsymbol{Y}}) = 0 \qquad \text{(Eq. 7.)}$$

Eq. 7. is generally solved using the Newton-Raphson method. Iterative estimates of $\boldsymbol{\beta}$ are obtained as:

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{Z} \qquad \text{(Eq. 8.)}$$

where $\boldsymbol{Z}$ is an $n \times 1$ column vector with elements:

$$z_i = \text{logit}[\hat{p}_i(\boldsymbol{x}_i)] + \frac{y_i - \hat{p}_i}{\hat{p}_i(1-\hat{p}_i)} \qquad \text{(Eq. 9.)}$$

and the weight matrix is:

http://www.ecology.uni-corvinus.hu ● ISSN 1589 1623

$$\boldsymbol{W} = \mathrm{diag}\left[\hat{p}_i(1-\hat{p}_i)\right] \qquad \text{(Eq. 10.)}$$

The covariance matrix of $\hat{\boldsymbol{\beta}}$ [1]:

$$Var(\hat{\boldsymbol{\beta}}) = \left\{\boldsymbol{X}^T \mathrm{diag}\left[p_i(1-p_i)\right]\boldsymbol{X}\right\}^{-1} \qquad \text{(Eq. 11.)}$$

Ridge regression to least squares estimation of linear models

The purpose of the parameter estimation is to find parameters as close to the true ones as possible. The most used parameter estimation methods lead to unbiased or asymptotically unbiased estimators. It means that the expected value of estimate is the true value of the parameter. However in some cases (e.g. when multicollinearity occurs) the unbiased estimators may have large variance which increases the probability of obtaining estimated parameters largely deviating from the true ones.

The goodness of an estimator is properly quantified by the mean square error function, which is defined for a scalar parameter as [4]:

$$\mathrm{MSE}\left(\hat{\beta}\right) = \mathrm{E}\left[\left(\hat{\beta}-\beta\right)^2\right] \qquad \text{(Eq. 12.)}$$

It is easy to show that MSE algebraically may be split into two parts:

$$\mathrm{MSE} = \mathrm{Var}\left(\hat{\beta}\right) + \left[\mathrm{bias}\left(\hat{\beta}\right)\right]^2 \qquad \text{(Eq. 13.)}$$

where

$$\mathrm{bias}\left(\hat{\beta}\right) = \mathrm{E}\left(\hat{\beta}\right) - \beta \qquad \text{(Eq. 14.)}$$

In multivariate case when $\boldsymbol{\beta}$ is a parameter vector, the mean square error function is defined as (in order to keep it scalar the trace of the covariance matrix is used):

$$\mathrm{MSE} = \mathrm{Tr}\left[\mathrm{Var}\left(\hat{\boldsymbol{\beta}}\right)\right] + \left[\mathrm{bias}\left(\hat{\boldsymbol{\beta}}\right)\right]^T \left[\mathrm{bias}\left(\hat{\boldsymbol{\beta}}\right)\right] \qquad \text{(Eq. 15.)}$$

In this sense a slightly biased estimator with smaller variance may be more advantageous than an unbiased estimator having large variance. Considering this Hoerl and Kennard (1970) have modified the least squares (LS) estimation for linear models and proposed a biased estimation method, called ridge regression [5].

Let us consider a linear model:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_h x_{ih} + \varepsilon_i \qquad \text{(Eq. 16.)}$$

where ε_i is the measurement error at the i-th point.

http://www.ecology.uni-corvinus.hu ● ISSN 1589 1623

In matrix notation

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon} \qquad \text{(Eq. 17.)}$$

where $\boldsymbol{Y}$ is an $n\times1$ vector of observable dependent variables, $\boldsymbol{\varepsilon}$ is an $n\times1$ vector of random errors.

The ordinary least squares (OLS) estimator is obtained by minimizing the following objective function:

$$\phi\left(\hat{\boldsymbol{\beta}}\right) = \left(\boldsymbol{Y} - \boldsymbol{X\hat{\beta}}\right)^{\mathrm{T}} \left(\boldsymbol{Y} - \boldsymbol{X\hat{\beta}}\right) = \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i} - \ldots - \hat{\beta}_h x_{hi}\right)^2 \qquad \text{(Eq. 18.)}$$

The estimated parameter vector is expressed as [6]:

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1} \boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y} \qquad \text{(Eq. 19.)}$$

The covariance matrix of $\hat{\boldsymbol{\beta}}$

$$\mathrm{Var}\left(\hat{\boldsymbol{\beta}}\right) = \sigma^2 \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1} \qquad \text{(Eq. 20.)}$$

Thus with Eq. 15.

$$\mathrm{MSE} = \sigma^2 \,\mathrm{Tr}\left[\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\right] = \sigma^2 \sum_{j=1}^{h} 1/\lambda_j \qquad \text{(Eq. 21.)}$$

where λ_j is the j-th eigenvalue of the $\boldsymbol{X}^T\boldsymbol{X}$ matrix.

If any of the λ_j eigenvalues is relatively small (it occurs when they differ in great extent) the value of MSE increases. It means that estimated parameters may be far from the true ones. Using ridge regression a small positive number is added to the diagonal elements of the $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ matrix:

$$\hat{\boldsymbol{\beta}}^* = \left(\boldsymbol{X}^T\boldsymbol{X} + k\boldsymbol{I}\right)^{-1} \boldsymbol{X}^T\boldsymbol{Y}, \qquad \text{(Eq. 22.)}$$

which may be also obtained by minimising the following objective function:

$$\phi\left(\hat{\boldsymbol{\beta}}^*\right) = \left(\boldsymbol{Y} - \boldsymbol{X\hat{\beta}}^*\right)^T \left(\boldsymbol{Y} - \boldsymbol{X\hat{\beta}}^*\right) + k\hat{\boldsymbol{\beta}}^{*T}\hat{\boldsymbol{\beta}}^* \qquad \text{(Eq. 23.)}$$

$\hat{\boldsymbol{\beta}}$ refers to ML and ridge estimators will be denoted by $\hat{\boldsymbol{\beta}}^*$. The ridge technique enlarges the small eigenvalue(s), thus decreases MSE. It is obvious that with k=0 the OLS estimator is recovered, while at $k\to\infty$ all $\hat{\beta}_j^*$ estimators go to zero.

The ridge estimator is proved to lead to smaller MSE than that obtained by the ordinary least squares method, if small enough positive value is chosen for k, that is

http://www.ecology.uni-corvinus.hu ● ISSN 1589 1623

$\mathrm{MSE}(\mathrm{OLS}) > \mathrm{MSE}(\mathrm{ridge})$.

While the existence of the minimum MSE is proved, the *k* value to which this optimum belongs to may not be calculated. Hoerl and Kennard (1970) proposed to use the ridge trace for deciding on *k*. The ridge trace is the plot of the estimated parameter values as function of *k*. When the $\hat{\beta}_j^*$ values cease to change strongly, the proper *k* is found.

The method is thoroughly discussed and applied in the literature [7-10], simulation studies were also performed.

Ridge method to logit regression

The MSE of asymptotically unbiased $\hat{\boldsymbol{\beta}}$ estimate with ML estimation from Eq. 11.:

$$\mathrm{MSE} = \mathrm{Tr}[\mathrm{Var}(\hat{\boldsymbol{\beta}})] = \mathrm{Tr}\left[\left\{\boldsymbol{X}^T \mathrm{diag}\left[p_i(1-p_i)\right]\boldsymbol{X}\right\}^{-1}\right] = \sum_{j=1}^{h} 1/\lambda_j \qquad \text{(Eq. 24.)}$$

where λ_j is the *j*-th eigenvalue of the $\boldsymbol{X}^T \mathrm{diag}\left[p_i(1-p_i)\right]\boldsymbol{X}$ matrix.

This is analogous to the variance of LS estimation (Eq. 20.). It is known that the eigenvalues of $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ differ in great extent if the columns of $\boldsymbol{X}$ matrix are correlated [2] (multicollinearity). This occurs when evaluating retrospective biomedical studies, where the regressor variables may not be set properly but may change in almost a random way. In Eq. 24. the estimated covariance matrix related to not simply $\boldsymbol{X}^T\boldsymbol{X}$ but $\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}$, where weights depend on $\boldsymbol{X}$ matrix. Thus even with orthogonal $\boldsymbol{X}$ matrix the eigenvalues of the $\boldsymbol{X}^T \mathrm{diag}\left[p_i(1-p_i)\right]\boldsymbol{X}$ matrix may differ. Ridge method is a remedial measure to treat multicollinearity with linear regression, and it can also be applied to the ML estimation as it was proposed by Schaefer [11]. A small positive number is added to the diagonal elements of the covariance matrix given by Eq. 11.:

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}^*) = \left\{\boldsymbol{X}^T \mathrm{diag}\left[n_i p_i(1-p_i)\right]\boldsymbol{X} + k\boldsymbol{I}\right\}^{-1} \qquad \text{(Eq. 25.)}$$

Thus the objective function has the form:

$$\phi(\hat{\boldsymbol{\beta}}^*) = \sum_{i=1}^{n} y_i \log(p_i) + \sum_{i=1}^{n} (1-y_i)\log(1-p_i) - k\hat{\boldsymbol{\beta}}^{*\mathrm{T}}\hat{\boldsymbol{\beta}}^* \rightarrow \max(\hat{\boldsymbol{\beta}}^*) \qquad \text{(Eq. 26.)}$$

the iterative estimate of parameter vector is obtained as:

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X} + k\boldsymbol{I}\right)^{-1}\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{Z} \qquad \text{(Eq. 27.)}$$

Elements of $\boldsymbol{Z}$ are defined by Eq. 9.

Barker and Brown [12] have compared ridge logit regression to principal component logistic regression and standard logistic regression by simulation studies. In [13] Cessie

http://www.ecology.uni-corvinus.hu ● ISSN 1589 1623

and Houwelingen applied the ridge logistic estimation to a biomedical problem. In this paper we apply the proposed technique both to continuous and discrete regressors, and investigate the effect of extent of correlation between regressors to the estimation error.

Example

We have analysed the data of Molnár M.Z. and co-workers [14]. In a part of their study they investigate the probability of occurrence of restless legs syndrome (RLS) for kidney-transplanted patients. We have fitted a logistic model to their data. The dependent variable is the prevalence of RLS, the covariates (three binomial and three continouos) are: diabetes, sex, modality (its value is 1 for the kidney-transplanted patients, and 0 for the waitlisted dialysis patients), age, albumin and haemoglobin (HGB) level. In this retrospective study the HGB level and modality are strongly correlated, thus the use of ridge regression seems to be recommended. The data set consists of data on 882 patients. (In the original study 992 patients were contained, parts of data for some of them missing. As our aim is to investigate the efficiency of logistic ridge estimators, and the problem of missing data analysis is not the scope of this paper, the cases with missing data were left out from the analysis.)

We have scaled the independent variable. The scaled variables change between 1 and 0. The aim of this transformation was twofold.

- The effect of the *j*-th covariate (on the dependent variable) depends on the range in which $x_j\beta_j$ changes: $effect(x_j) = range(x_j)\beta_j = (x_{j,\max} - x_{j,\min})\beta_j$. If $range(x_j)$ is equal for each *j*, the value of β_j indicates the importance of the *j*-th covariate.
- The model estimates the effect of the *j*-th covariate. The error of this estimated effect is related to $range(x_j)s_{\beta_j}$, where s_{β_j} is the standard deviation of β_j. If $range(x_j)$ is equal for each *j*, s_{β_j} measures the error of the effect of the *j*-th covariate. In the following only the scaled x_j covariates and the β_j scaled model parameters will be used.

Using the data of the 882 patients the $\boldsymbol{\beta}$ parameter vector was estimated both with ML- and with ridge logistic regression. The obtained estimated model parameters do not differ in great extent for the two estimation procedures. Using the bootstrap method [3] a 95% bootstrap interval was calculated for each β_j. The results are shown on *Fig. 1*.

http://www.ecology.uni-corvinus.hu ● ISSN 1589 1623

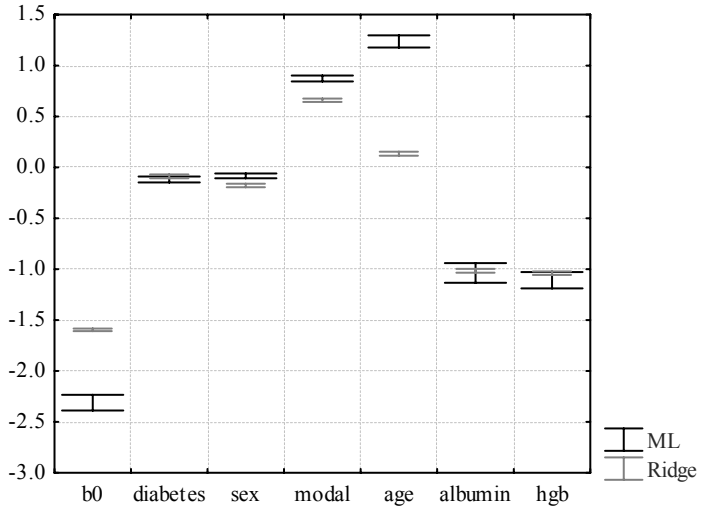

Figure 1. *95% bootstrap intervals for the estimated model parameters with ML and with ridge estimation for the full (882 patients) data set (shorter whiskers belongs to the ridge estimation)*

It can be seen from the figure that 1) the variance of the ridge estimation is smaller than that of the ML estimation, but they are approximately of the same order of magnitude; 2) the means of the parameters estimated by the two regression methods differ considerably, the ridge estimators are shrunken toward 0. In this case the use of ridge regression is not reasonable, because the smaller variance of the ridge estimator do not compensate its bias. Due to the large sample size the variance of ML estimator is relatively small, the use of ridge estimation is not justified. The advantages of using the ridge regression in case of smaller sample sizes may still be a relevant question. This is the scope of this study.

As the full data set is large enough, the ML estimated model parameters from it are close enough to the true model parameters, thus they will be considered as such. These parameters are used to evaluate MSE in further calculations.

We have obtained samples of size n, with replacement, from the original (N=882) dataset. (n:100, 110, 120, 125, 130, 140, 180, 200, 300, 400, 600) Having the sample size fixed, m_n samples were obtained (e.g. for n=100, m_{100} =800 sample were taken, each of size n). We have fitted the logistic model for each sample both with ML and with ridge regression. The estimation error (MSE) is calculated using Eq 28. for each n for both estimation procedures.

$$\mathrm{MSE}(\hat{\boldsymbol{\beta}})_n = \frac{\sum_{i=1}^{m_n}\sum_{j=1}^{h}\left(\hat{\beta}_{j,i} - \beta_j\right)^2}{m_n} \qquad \text{(Eq. 28.)}$$

The results are show in *Fig. 2*.

http://www.ecology.uni-corvinus.hu ● ISSN 1589 1623

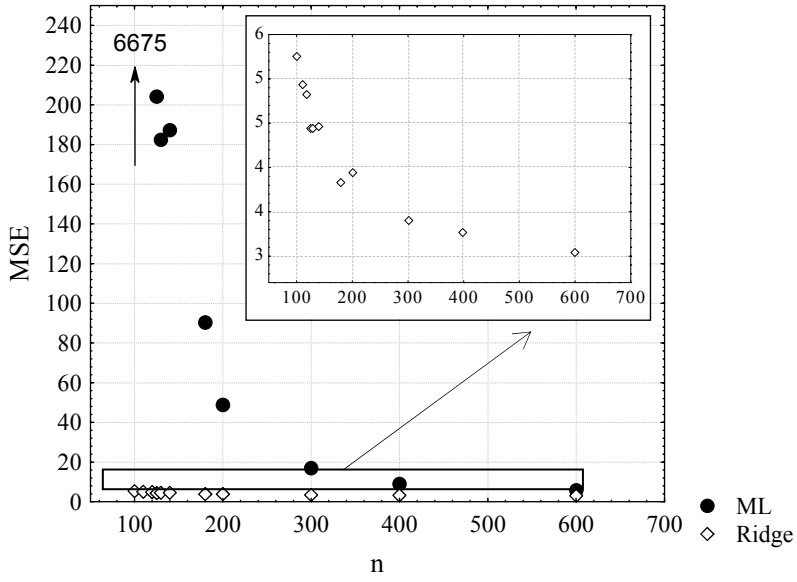

Figure 2. *Error of estimated model parameters with ML and ridge estimation, respectively versus sample size*

The error of the parameter vector ($\mathrm{MSE}(\hat{\boldsymbol{\beta}})$) of ML estimators is exponentially increasing as the sample size is decreasing. The error of ridge estimator changes similarly (*Fig. 2*), but its variation is by orders of magnitude smaller. The use of ridge estimation seems to be useful if the sample size is low, for this example if it is less than about 300.

Conclusions

We have compared the effectiveness of logistic ridge and ML regression using clinical data of kidney-transplanted patients. The use of ridge regression is not recommended for large samples. In these cases the variance of ML estimation is relatively small, thus the variance reduction achieved with ridge estimation dos not compensate the bias of the method. For smaller sample sizes the variance of ML estimator increases strongly as the sample size decreasing, while the variance of ridge estimation hardly changes. Thus for smaller samples the use of ridge method proved to be more effective then the ML estimation.

Acknowledgement. The authors acknowledge Quality of Life Research Team of Institute of Behavioural Sciences, Semmelweis University for providing the data.

REFERENCES

[1] A. Agresti (2002): Categorical data analysis. – John Wiley & Sons, New York, p. 166., 194.

http://www.ecology.uni-corvinus.hu ● ISSN 1589 1623

[2] R.H. Myers (1989): Classical and modern regression with applications. – PWS-KENT, Boston, , p.318., p.126.
[3] T.P. Ryan (1997): Modern regression methods. – John Wiley & Sons, New York, p. 260.
[4] S.H. Ngo, S. Kemény, A. Deák (2003): – Chemometrics and Intelligent Laboratory Systems 67 69-78
[5] A.E. Hoerl and R. W. Kennard (1970): . – Technometrics 12 55-67.
[6] N.R. Draper, H. Smith (1998): . – Applied regression analysis, John Wiley & Sons, New York, , p.136.
[7] J.W. Gorman, R.J. Toman (1966): . – Technometrics 8 27-51.
[8] R.F. Gunst , R.L. Mason (1977): . – Biometrics 33 249-260.
[9] H. Pasternak, Z. Schmilovitch, E. Fallik, Y. Edan (2001), . – Journal of Testing and Evaluation 29 60-66.
[10] J.H. Kalivas (2001): . – Analitica Chimica Acta 428 31-40.
[11] RL. Schaefer (1986): . – Journal of Statistical Computation and Simulation 25 75-91
[12] L. Barker, C.Brown (2001): . – Statistics in Medicine 20 1431-1442
[13] S. Cessie J.C. Houwelingen (1992): . – Applied Statistics 41 No.1191-201
[14] M.Z. Molnár, Novak M., Ambrus C., Szeifert L., Kovacs A., Pap J., Remport A., I. Mucsi (2005): . – Am J Kidney Dis 45(2) 388-96.

http://www.ecology.uni-corvinus.hu ● ISSN 1589 1623