Journal of Data Science **6**(2008), 515-531

A Solution to Separation and Multicollinearity in Multiple Logistic Regression

Jianzhao Shen and Sujuan Gao
Indiana University School of Medicine

Abstract: In dementia screening tests, item selection for shortening an existing screening test can be achieved using multiple logistic regression. However, maximum likelihood estimates for such logistic regression models often experience serious bias or even non-existence because of separation and multicollinearity problems resulting from a large number of highly correlated items. Firth (1993, *Biometrika*, **80**(1),27-38) proposed a penalized likelihood estimator for generalized linear models and it was shown to reduce bias and the non-existence problems. The ridge regression has been used in logistic regression to stabilize the estimates in cases of multicollinearity. However, neither solves the problems for each other. In this paper, we propose a double penalized maximum likelihood estimator combining Firth's penalized likelihood equation with a ridge parameter. We present a simulation study evaluating the empirical performance of the double penalized likelihood estimator in small to moderate sample sizes. We demonstrate the proposed approach using a current screening data from a community-based dementia study.

Key words: Logistic regression, maximum likelihood, penalized maximum likelihood, ridge regression, item selection.

1. Introduction

In dementia studies involving large community-based cohorts of elderly participants, screening tests are often administered to study participants in order to obtain a measure of cognitive function. The scores of the screening tests are then used to determine whether a study participant should undergo more intensive clinical assessment for diagnosis of dementia. Such screening tests are usually constructed using a number of items which aim at measuring various domains of cognitive function. In practical epidemiological or clinical research, however, there is always a need to shorten existing screening instruments in order to more efficiently gather the needed information in a limited amount patient time.

The selection of test items for such shortened screening tests has often involved univariate comparisons of percentage of correct responses between demented and normal subjects. Items are retained if they show significant difference in responses between the two groups of subjects. However, the univariate method only considers each individual item and ignores that responses for multiple items are often highly correlated. Combining items with the largest differences in item responses between demented and normal groups may not lead to the most discriminating overall test.

A natural alternative to the univariate method would be multivariate item selection using a multiple logistic regression model, which multivariately models the relationship between dementia outcomes and the screening items. The multiple logistic regression model serves several purposes. One is that the model offers the collective predictive accuracy of the items. The second appeal is that the model also provides a linear combination of all test items which may be used as a score to predict the dementia outcomes. Lastly, the standardized parameter estimates from the logistic model also offer the rankings of test items in terms of strength of association with dementia.

Item selection using multiple logistic regression often encounters serious estimation problems when applied to screening data in dementia. Separation and multicollinearity are the two common problems in the logistic regression. The problems become exasperated in the dementia screening data because the two problems frequently occur together. These problems in logistic regression have led to aborted attempts by many investigators to shorten instruments for screening purposes using multiple logistic regression.

Separation in logistic regression frequently occurs when the binary outcome variable can be perfectly separated by a single covariate or by a non-trivial linear combination of the covariates (Albert and Anderson (1984)). Heinze and Schemper (2002) demonstrated that separation can happen even when the underlying model parameters are low in absolute value. They also showed that the probability of separation depends on sample size, on the number of dichotomous covariates, the magnitude of the odds ratios and the degree of balance in their distribution. Ceiling effect resulting from a high proportion of sample having maximum score in item response also increases the probability of separation.

Separation alone can cause infinite estimates or biased estimates. In cases where finite maximum likelihood estimates are available, correction of bias in maximum likelihood estimates of logistic regression have been studied by Anderson and Richardson (1979), McLachlan (1980), Schaefer (1983), Copas (1988), McCullagh and Nelder (1989) and Cordeiro and McCullagh (1991). In situations where finite maximum likelihood estimates may not exist, Firth (1992, 1993) introduced a penalized maximum likelihood estimator which reduced the bias of

maximum likelihood estimates in generalized linear models. Firth's approach guarantees the existence of estimates by removing the first order bias at each iteration step. Bull et al. (2002) extended Firth's approach to multinomial logistic regression with nominal response categories, comparing it to maximum likelihood estimates and to maximum likelihood estimates corrected by an estimate of the asymptotic bias. They showed that Firth's penalized maximum likelihood estimator was superior to the other methods in small samples and Firth's estimator was equivalent to the maximum likelihood estimator as sample size increased. Heinze and Schemper (2002) applied Firth's approach to logistic regression with separated data sets. Their simulation results demonstrated that Firth's penalized likelihood estimator provided an ideal solution to the separation problem in logistic regression.

Multicollinearity is not uncommon when there is a large number of covariates. It may become a serious concern in dementia data because screening items are often highly correlated. Multicollinearity can cause unstable estimates and inaccurate variances which affects confidence intervals and hypothesis tests (Hoerl and Kennard, 1970, 1988). A ridge estimator originally developed for linear regression provides a way to deal with the problems caused by multicollinearity. The ridge estimator in general shrinks estimates towards the origin. The amount of shrinkage is controlled by the ridge parameter, whose size depends on the number of covariates and the magnitude of collinearity. The mean squared error (MSE) is guaranteed to be reduced accordingly by the introduction of ridge parameter (Schaefer et.al, 1984; Hoerl and Kennard,1970,1988). le Cessie and van Houwellingen (1992) applied the ridge regression method to logistic regression to improve parameter estimates and decrease prediction errors.

In this paper, we propose a double penalized maximum likelihood estimator for logistic regression models which combines Firth's penalized likelihood approach with a second penalty term for ridge parameter which is capable of handling both separation and multicollinearity. In section 2, we describe the proposed approach in detail. In section 3, we evaluate the empirical performance of the proposed estimator in a simulation study. In section 4, We demonstrate the proposed approach using screening item data from a community-based dementia study.

2. Proposed Approach

2.1 Double penalized likelihood estimator

Let Y_i be a binary response variable, $i = 1, \ldots, n$. Let $x_i = (x_{i1}, \ldots, x_{ip})$ be a p-dimentional row vector of covariates. We denote the covariate matrix by X. In a logistic model, we model the probability $\mathrm{P}(Y_i = 1)$ by

$$P(Y_i = 1) = \pi_i = \frac{1}{1 + \exp^{-x_i\beta}}$$

where β is a $p \times 1$ vector of parameters corresponding to the covariates. The log-likelihood function under the logistic regression model is given by

$$L(\beta) = \sum_{i=1}^{n} l_i(\beta),$$

where

$$l_i(\beta) = Y_i \log \pi_i + (1 - Y_i) \log(1 - \pi_i). \tag{2.1}$$

Maximizing $L(\beta)$ yields the maximum likelihood estimate (MLE) of β.

Maximum likelihood estimates of β are found to be biased away from the point $\beta = 0$. The asymptotic bias of the MLE $\hat{\beta}$ can be expressed as:

$$\text{Bias}(\hat{\beta}) = \frac{b_1(\beta)}{n} + \frac{b_2(\beta)}{n^2} + \dots.$$

Most bias correction methods are focused on removing the first-order bias from the asymptotic bias of $\hat{\beta}$ by using:

$$\hat{\beta}_{\text{corrected}} = \hat{\beta} - \frac{b_1(\hat{\beta})}{n}.$$

However, this method relies on the existence of the MLE $\hat{\beta}$. In small to medium sized data, or data with many covariates, it is not uncommon for $\hat{\beta}$ to be infinite. Let A be the Fisher information matrix for $l(\beta)$. Firth (1993) proposed the following penalized likelihood function:

$$L^*(\beta) = L(\beta) A^{1/2}$$

with the penalized log-likelihood function of

$$l^*(\beta) = l(\beta) + \frac{1}{2} \log |A|. \tag{2.2}$$

Firth (1992) noted that this penalized likelihood function is equivalent to the posterior distribution function for a canonical parameter of an exponential family model using the Jeffrey's invariant prior (1946).

We propose to add a second penalty term to Firth's penalized likelihood function by including a ridge parameter which forces the parameters to spherical restrictions:

$$l^{**}(\beta) = l(\beta) + \frac{1}{2} \log |A| - \lambda \parallel P\beta \parallel^2, \tag{2.3}$$

where P is a $p \times p$ matrix which imposes linear constrains on the parameters of β and $\| \|$ denotes the Euclidean norm of a parameter vector. When P is the identity matrix, all parameters in β equally shrink towards the origin. A special situation for P would be a partial diagonal matrix with 1 at some diagonal elements and 0 elsewhere and the parameters corresponding to 0 will not be subjected to any shrinkage. Without loss of generalization, we consider P to be an identity matrix in subsequent development.

In the double penalized likelihood function $l^{**}(\beta)$, λ is the ridge parameter which controls the amount of shrinkage of the norm of β. The choice of λ depends on the number of covariates and/or the degree of multicollinearity among the covariates.

In a simple and intuitive example, suppose we have just one independent variable, x_i, which can take only two values: 0 or 1. We consider the simplest logistic regression of

$$P(Y_i = 1) = \frac{1}{1 + e^{-x_i\beta}}.$$

In other words, we consider a simple logistic regression with just one regression parameter without an intercept. Now let $n_{11} = \sum(y_i = 1|x_i = 1)$, $n_{01} = \sum(y_i = 0|x_i = 1)$, $n_1 = n_{11} + n_{01}$, $n_{10} = \sum(y_i = 1|x_i = 0)$ and $n_{00} = \sum(y_i = 0|x_i = 0)$. In addition, let $p_1 = \text{Prob}(y_i = 1|x_i = 1) = \frac{1}{1+e^{-\beta}}$. Since $p_0 = \text{Prob}(y_i = 1|x_i = 0) = 1/2$, the likelihood function can be written as:

$$L = p_0^{n_{10}}(1 - p_0)^{n_{00}} p_1^{n_{11}}(1 - p_1)^{n_1 - n_{11}}.$$

Now let $C = p_0^{n_{10}}(1 - p_0)^{n_{00}}$ and it can be seen that C is a constant not involving the parameter β. Hence the corresponding log-likelihood function can be written as:

$$l = \log C + n_{11}\beta + n_1 \log(1 - p_1).$$

The maximum likelihood estimate of β for this model is:

$$\hat{\beta} = \log \frac{n_{11}}{n_1 - n_{11}}.$$

It is clear from the formulae above that when $n_1 = n_{11}$, meaning that there is no observation with the combination of $x_i = 1$ and $y_i = 0$, a complete separation of data will happen. Under such data separation, the maximum likelihood estimate for β is infinite, i.e. $\hat{\beta} = \infty$.

For this simple logistic model, Fisher's information matrix is reduced to $A = n_1 p_1(1 - p_1)$. Thus Firth's penalized likelihood function for this model is written as:

$$L^* = C p_1^{n_{11} + \frac{1}{2}}(1 - p_1)^{n_1 - n_{11} + \frac{1}{2}}.$$

Parameter estimate from Firth's penalized likelihood function is therefore:

$$\hat{\beta}^* = \log\frac{n_{11} + \frac{1}{2}}{n_1 - n_{11} + \frac{1}{2}}.$$

Notice that Firth's penalized likelihood estimate is essentially the frequently used method of adding $\frac{1}{2}$ to a zero cell.

Our proposed double penalized log-likelihood function is:

$$l = \log C + (n_{11} + \frac{1}{2})\beta + (n_1 + 1)\log(1 - p_1) - 2\lambda\beta.$$

Parameter estimate from the double penalized likelihood approach for this simple logistic model becomes:

$$\hat{\beta}^{**} = \log\frac{n_{11} + \delta}{n_1 - n_{11} + (1 - \delta)},$$

where $\delta = \frac{1}{2} - \lambda$. It can be seen that when $\lambda = 0$, $\delta = \frac{1}{2}$, thus the double penalized likelihood approach is the same as Firth's penalized approach. From this sense, our proposed double penalized approach is an extension to Firth's method with additional ability to choose λ to reduce multicollinearity. See Section 2.3 below for further discussion on the choice of λ.

2.2 Parameter estimation

The Newton-Raphson maximization algorithm procedure can be used to derive $\hat{\beta}$ for β. The iterative Newton-Raphson algorithm is defined as:

$$\hat{\beta}^{(t+1)} = \hat{\beta}^{(t)} + (A^{**})^{-1(t)} U^{**}(\hat{\beta}^{(t)}),$$

where t stands for the number of iteration, $(A^{**})^{-1(t)}$ is the information matrix of the double penalized likelihood function in (2.3), and $U^{**}(\hat{\beta}^{(t)})$ is the first derivative of the log-likelihood function, i.e.

$$U^{**}(\beta) = \frac{\partial l(\beta)}{\partial \beta} + \frac{\partial}{\partial \beta}(\frac{1}{2}\log|A|) - 2\lambda\beta.$$

A computationally convenient formulae in matrix form for $\frac{\partial}{\partial \beta}(\frac{1}{2}\log|A|)$ was provided by Bull *et al* (2002) as:

$$\frac{\partial}{\partial \beta}(\frac{1}{2}\log|A|) = X'Q(X \otimes X)vec(A^{-1}),$$

where $Q = W \otimes e_i$, $i = 1, \ldots, n$, W is a $n \times n$ diagonal matrix with element $\pi_i(1 - \pi_i)(1 - 2\pi_i)$, e_i is a $1 \times n$ vector with 1 on the ith column and 0 elsewhere, $vec(A^{-1})$ is the vectorized A^{-1}.

Since the information matrix for the double penalized likelihood function, $A^{**} = -\frac{\partial U(\beta)}{\partial \beta} + \frac{1}{2}\frac{\partial^2 \log|A|}{\partial \beta^2} - 2\lambda I$, is difficult to derive, we propose to use

$$\tilde{A}^{**} \approx -\frac{\partial U(\beta)}{\partial \beta} - 2\lambda I.$$

In Bull *et al.* (2002) where Firth's original penalized likelihood was considered, the authors proposed to use the information matrix from the likelihood function which also excluded Firth's penality term. The approximate information matrix can also be used to provide variance estimates of the parameters at convergence. In a simulation study presented in the following section, we evaluate the approximate variance estimates using empirical variance estimates.

2.3 The choice of the ridge parameter

The ridge parameter,λ,is generally chosen to minimize prediction errors of the logistic models. There are various definitions of the prediction errors in the literature. Efron(1986), le Cessie and van Houwelingen (1992), for example, defined three measures to quantify the prediction errors:

(1) classification error (CE)

$$\text{CE} = \begin{cases} 1 & \text{if } Y_{\text{new}} = 1 \text{ and } \hat{\pi} < \frac{1}{2}; \text{ or } Y_{\text{new}} = 0 \text{ and } \hat{\pi} > \frac{1}{2}, \\ \frac{1}{2} & \text{if } \hat{\pi} = \frac{1}{2}, \\ 0 & \text{otherwise;} \end{cases}$$

(2) squared error (SE)

$$SE = (Y_{new} - \hat{\pi})^2;$$

(3) minus log-likelihood error (ML)

$$ML = -\{Y_{new}\log\hat{\pi} + (1 - Y_{new})\log(1 - \hat{\pi})\}.$$

The advantages and disadvantages of these methods were reviewed in le Cessie and van Houwelingen (1992). Unlike the classification error approach which considers model prediction in the neighborhood of $\pi = \frac{1}{2}$, the squared error approach and minus log-likelihood error approach calculate the prediction errors in the entire range of π values. In this paper, we focus the squared error as a measure of prediction errors because of its computational convenience.

The mean squared error (MSE) is calculated as an index based on the entire data set when comparing the effects of different ridge parameters. In lack of an external validation dataset, we used cross-validated estimate of the mean squared error (MSE) which is defined as:

$$\text{MSEcv}^{\lambda} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{\pi}_i^{\lambda})^2,$$

where $\hat{\pi}_i^{\lambda}$ denotes the prediction of π_i at a given λ based on parameter estimates obtained with the ith observation left out. The optimal λ was chosen to minimize the MSE.

For ease of computation in the cross-validation, we adopted a one-step approximation method proposed by Cook and Weisberg (1982) for the estimation of $\beta_{(-i)}^{\lambda}$, the parameters with i^{th} observation left out:

$$\hat{\beta}_{(-i)}^{\lambda} = \hat{\beta}^{\lambda} - \frac{\{\tilde{A}^{**}\}^{-1} X_i^T (Y_i - \hat{\pi}_i^{\lambda})}{1 - h_{ii}},$$

where h_{ii} denotes the i^{th} diagnonal element in the hat matrix generated from $\tilde{A}^{**}$. Similar to le Cessie and van Houwelingen (1992) MSE from cross-validation can be approximated by:

$$\text{MSEcv}^{\lambda} \approx \frac{1}{n} \sum_{i=1}^{n} \frac{(Y_i - \hat{\pi}_i^{\lambda})^2}{(1 - h_{ii})^2}.$$

The asymptotic consistency of the double penalized likelihood estimator was shown in Gao and Shen(in revision). We investigate the empirical performances of the proposed estimator in small to moderate samples in the following section.

3. A Simulation Study

A simulation study was conducted to evaluate the empirical performance of the double penalized likelihood estimator (DPLE) in logistic regression and compared it to the maximum likelihood estimator (MLE) and Firth's penalized likelihood estimator (PLE) with respect to mean bias and mean squared error (MSE). We considered various scenarios where there were a large number of covariates with potential high correlations. A combination of binary and continuous covariates was included in the simulation. We chose ten covariates with five binary and five continuous covariates in the simulation so that the demand of the large number of covariates was met. The sample sizes were set to be 30, 50, 80, 130 and 200 so that we could investigate how the performance changes with the change of sample sizes. We implemented the algorithm using SAS IML and macro languages (SAS 8.2). A SAS macro is available upon request.

We first generated 10 continuous variates from multivariate normal distributions with zero means, unit variances and a specified correlation matrix, followed by dichotomization at zero for the first five continuous variates to produce the five binary covariates. The correlation coefficients among the variates from the multivariate normal distributions were set to be equally as high as 0.8. The binary response variable was generated by comparing the predicted probabilities

from the logistic model with true parameter values to a random number from the uniform distribution in $[0, 1]$.

We simulated 1000 datasets for each sample size scenario . Our results confirmed that the proportion of non-existence for the regular MLEs were extremely high in small samples but decreased as sample size increased. In our simulations, none of the datasets had finite MLEs when the sample size was 30. The proportion of having finite MLEs increased as sample size increased (5%, 33%, 83% and 100% for sample size 50, 80, 130 and 200 respectively). Results of the simulations are presented in Table 1.

Table 1: Simulation results for logistic regression with 10 covariates using three estimation methods

Sample Size	Parameter	Mean Bias			Mean Squared Error			% Bias(Variance)	
(% Finite MLE)	(True)*	MLE	PLE	DPLE	MLE	PLE	DPLE	PLE	DPLE
$n = 30$	β_0 (-2)		0.19	0.36		1.19	1.23	301	322
(0%)	β_1 (1)		-0.25	-0.31		1.46	1.27	392	473
	β_2 (1)		-0.22	-0.29		1.35	1.19	662	796
	β_3 (1)		-0.24	-0.31		1.25	1.12	450	526
	β_4 (1)		-0.21	-0.26		1.41	1.19	436	534
	β_5 (1)		-0.26	-0.33		1.44	1.29	408	473
	β_6 (1)		-0.86	-0.84		2.21	2.04	341	390
	β_7 (1)		-0.90	-0.88		2.32	2.12	339	381
	β_8 (1)		-0.84	-0.82		2.09	1.91	322	367
	β_9 (1)		-0.88	-0.85		2.25	2.05	306	344
	β_{10} (1)		-0.87	-0.85		2.26	2.08	303	331
$n = 50$	β_0 (-2)	-0.62	-0.04	0.31	12.45	1.87	1.51	156	164
(5%)	β_1 (1)	0.64	-0.15	-0.27	9.80	1.45	1.09	153	187
	β_2 (1)	0.16	-0.04	-0.19	17.57	1.49	1.14	159	188
	β_3 (1)	-0.14	-0.14	-0.26	8.73	1.65	1.30	126	148
	β_4 (1)	0.11	-0.06	-0.21	6.50	1.48	1.09	142	175
	β_5 (1)	1.08	-0.03	-0.18	7.89	1.54	1.13	137	158
	β_6 (1)	0.54	-0.61	-0.59	7.12	1.96	1.57	175	213
	β_7 (1)	1.34	-0.68	-0.64	17.71	2.11	1.65	162	178
	β_8 (1)	0.27	-0.63	-0.61	6.59	2.01	1.65	179	219
	β_9 (1)	0.05	-0.61	-0.59	12.49	1.96	1.55	155	193
	β_{10}(1)	0.30	-0.67	-0.63	15.23	2.16	1.67	165	206
$n = 80$	β_0 (-2)	-1.67	0.00	0.36	37.96	2.19	1.51	64	84
(33%)	β_1 (1)	0.77	-0.10	-0.23	18.07	1.60	1.09	61	85
	β_2 (1)	0.49	-0.05	-0.20	8.68	1.33	0.92	87	116
	β_3 (1)	0.99	-0.02	-0.17	26.22	1.42	0.97	78	99
	β_4 (1)	1.01	-0.04	-0.19	25.90	1.47	1.05	67	84
	β_5 (1)	0.85	-0.07	-0.23	22.09	1.40	0.94	88	125
	β_6 (1)	0.86	-0.30	-0.32	18.94	1.78	1.19	82	113
	β_7 (1)	1.14	-0.23	-0.28	17.64	1.74	1.08	78	115
	β_8 (1)	1.02	-0.28	-0.30	20.52	1.63	1.06	81	108
	β_9 (1)	1.06	-0.28	-0.31	23.00	1.73	1.15	91	127
	β_{10} (1)	0.90	-0.31	-0.34	21.41	1.72	1.17	79	102

*Parameter β_0 denotes the intercept, β_1-β_5 denote binary covariates and β_6-β_{10} denote continuous covariates.

Table 1 (continued): Simulation results for logistic regression with 10 covariates using three estimation methods

Sample Size	Parameter	Mean Bias			Mean Squared Error			% Bias(Variance)	
(% Finite MLE)	(True)*	MLE	PLE	DPLE	MLE	PLE	DPLE	PLE	DPLE
$n = 130$	β_0 (-2)	-1.21	0.01	0.28	16.30	1.35	0.95	30	52
(83%)	β_1 (1)	0.62	-0.06	-0.17	8.05	1.03	0.74	24	42
	β_2 (1)	0.49	-0.04	-0.15	10.45	1.17	0.83	15	29
	β_3 (1)	0.62	-0.02	-0.13	8.08	1.01	0.70	28	46
	β_4 (1)	0.83	0.04	-0.09	10.09	1.00	0.70	28	44
	β_5 (1)	0.78	-0.01	-0.12	10.47	1.10	0.75	18	36
	β_6 (1)	0.80	-0.06	-0.13	11.71	1.25	0.82	25	42
	β_7 (1)	0.89	-0.08	-0.14	10.50	1.20	0.83	35	54
	β_8 (1)	1.03	-0.04	-0.10	15.44	1.11	0.74	36	60
	β_9 (1)	0.89	-0.03	-0.10	14.57	1.17	0.76	35	60
	β_{10} (1)	0.82	-0.06	-0.12	10.32	1.13	0.76	22	43
$n = 200$	β_0 (-2)	-0.72	0.00	0.17	5.38	0.89	0.65	-1	20
(100%)	β_1 (1)	0.34	-0.04	-0.11	2.81	0.69	0.53	-3	11
	β_2 (1)	0.44	0.03	-0.05	2.83	0.61	0.47	6	22
	β_3 (1)	0.40	0.03	-0.05	2.05	0.61	0.47	5	21
	β_4 (1)	0.36	-0.01	-0.09	2.40	0.60	0.46	8	26
	β_5 (1)	0.38	0.01	-0.07	1.95	0.60	0.46	7	23
	β_6 (1)	0.47	0.00	-0.03	3.25	0.69	0.53	6	20
	β_7 (1)	0.46	0.03	-0.02	2.40	0.72	0.55	1	16
	β_8 (1)	0.52	0.04	-0.01	3.14	0.76	0.56	3	15
	β_9 (1)	0.47	0.00	-0.05	2.90	0.75	0.57	9	21
	β_{10} (1)	0.46	0.01	-0.03	2.41	0.70	0.53	7	23

*Parameter β_0 denotes the intercept, β_1-β_5 denote binary covariates and β_6-β_{10} denote continuous covariates.

In general, both penalized maximum likelihood estimators had significantly smaller mean biases and MSEs compared to the MLEs. The differences are larger when the sample size is small or medium (see Table 1). In Figure 1, we plotted the mean bias and the MSE of the PLE and DPLE for different sample sizes. It can be seen that DPLE has bigger bias than PLE and the MSE of DPLE was smaller compared to that of PLE. This trend is more evident when the sample size is small or moderate. As the sample size increased, the two procedures became indistinguishable with respect to the mean bias and MSE. We point out that these results are expected because the introduction of the ridge parameter in the DPLE sacrifices bias for the gain in MSE.

There seem to be differences in the parameter estimates for the binary and continuous covariates between the PLE and DPLE estimators with respect to mean bias and MSE in our simulation. The ridge effect appeared to be stronger in continuous covariates than in binary covariates. For the binary covariates, the reduction of MSE in DPLE compared to PLE clearly showed a compromise in terms of increased bias. For the continuous covariates, the MSE was significantly reduced in the DPLE method. However, the mean bias was not significantly affected by the addition of the ridge parameter. DPLE and PLE differed most

notably in smaller sample sizes and their differences become negligible when the sample size gets large.

We also included the percentage of estimated bias using the approximate variance estimates compared to empirical variance estimates based on the simulation in Table 1. The approximate information matrix overestimated the variances in both DPLE and PLE estimators when the sample size was small. The magnitude of overestimation is significantly reduced as sample size increased.

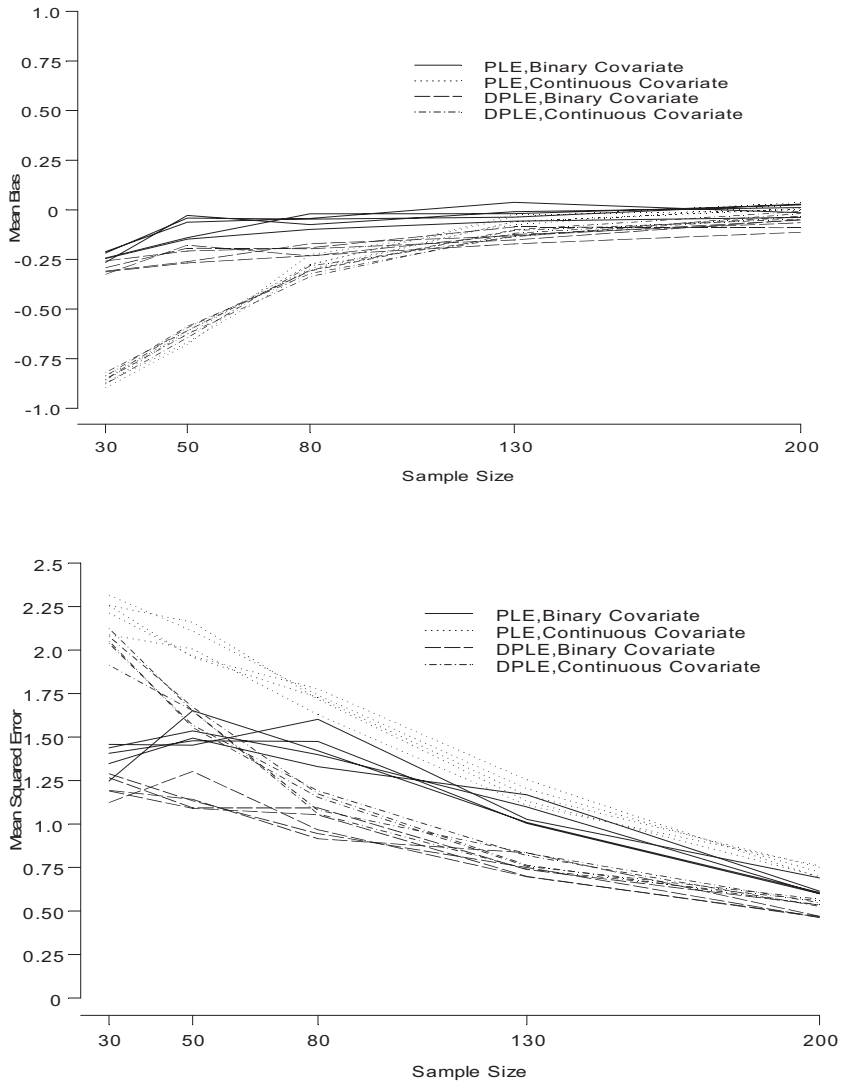

Figure 1: Plots of mean biases and mean squared errors for sample sizes and covariate types in logistic regression

4. Indianapolis Dementia Study

The Indianapolis Dementia Study is a community-based prospective cohort study of Alzheimer's disease and dementia in elderly African Americans in Indianapolis, Indiana. Data was collected during one baseline wave and four follow-up waves at 2, 5, 8 an 11 years after baseline. An enrichment sample of additional subjects was added at the third follow-up wave when blood samples were also collected from those subjects who consented for biochemical analyses. The Community Screening Interview for Dementia (CSID) was used as a screening instrument to evaluate and stratify study subjects' cognitive function (Hall *et al.*, 1996). Items in the CSID were selected from various well-known neuropsychological tests to measure the following functions in an interview: memory, abstract thinking, reasoning and judgement. There are 33 test questions in the CSID. At the third follow-up examination, 2628 elderly African-American subjects were screened with the CSID and stratified into three performance groups, namely poor, intermediate and good performance. Four hundred and fifty eight subjects were randomly selected with the sampling probabilities of 100%, 50% and 5%, respectively, in the poor, intermediate and good performance groups to undergo extensive clinical evaluation for dementia and Alzheimer's disease.

In this analysis, we selected 5 binary and 5 continuous items from the CSID which were presumed predictive of dementia. We included subjects who were administered the CSID for the first time at the 8-year wave and who also had biochemical measures available. Furthermore, we excluded subjects with a diagnosis of mild cognitive impairment and subjects with incomplete responses to screening items, resulting in a sample of 59 subjects, of whom 21 were demented and 38 were normal. The items are described in details as below:

1. Do you remember my name? (interviewers introduced themselves at the beginning of interview and asked the subject to remember their names for later recall)
2. Who is the Mayor of the city?
3. What is the name of the civil rights leader who was assassinated in Memphis in 1968?
4. Who is the current president in the USA?
5. Who is the current governor of Indiana?
6. Place orientation. It is generated from the following 4 binary items:
   1) What is the name of the city?
   2) What are the two major streets near your home?
   3) What is the city market?
   4) What is your complete address, including your zip code?
7. Time orientation. it is generated from the following 5 binary items:

1) What day of the week is it?
2) What month is it?
3) What year is this?
4) What season is it?
5) Did it rain(snow)yesterday?

8. Read a short story and then ask the subject to repeat as much of the story as he/she can.
9. Recall the story from item 8 after a while.
10. Ask the subject to give the names for as many different animals as he/she can in one minute.

Table 2: Demographic characteristics and item properties in the demented and normal groups

Variables	Demented ($n = 21$)	Normal ($n = 38$)	p-value*
Demographic Characteristics			
Mean age (SD)	79(5.7)	77(5.3)	0.1097
Mean years of education (SD)	8.6(4.4)	9.6(3.3)	0.3373
% Male	62	40	0.0985
% White Collar Occupation**	15.6	24.8	0.1202
Binary Items	%correct	%correct	
Remember Name	33	89	< 0.0001
Mayor	10	32	0.0428
Civil War	67	92	0.0156
President	24	87	< 0.0001
Governor	14	32	0.1441
Continuous Items	Mean(SD)	Mean(SD)	
Place Orientation	2.5(1.3)	3.8(0.5)	0.0002
Time Orientation	2.9(1.5)	4.7(0.5)	< 0.0001
Repeat Story	2.8(1.8)	4.9(1.3)	< 0.0001
Recall Story	2.5(2.0)	4.7(1.6)	< 0.0001
Animals	8.7(3.9)	14.1(4.0)	< 0.0001

*p-value is obtained from t-tests for continuous variables and from χ^2 tests for binary variables.
**White Collar Occupations included sales, professional, clerical and/or administrative, and protective. Occupations were classified using the Standardized Occupational Classification Manual.

The first 5 items are binary with correct and incorrect responses and the last 5 items are continuous. The items "Place orientation" and "Time orientation" are highly correlated (r = 0.74). Table 2 presents demographic characteristics and

item responses between the demented and normal groups. All items except "Governor" have significantly different responses univariately between the demented and normal.

The logistic regression models using the MLE, PLE and DPLE are presented in Table 3. Figure 2 demonstrated that the cross-validated estimate of the MSE changes as a function of the ridge parameter. The optimal ridge parameter is chosen to be 0.01 for this data set. Table 3 showed the two penalized MLEs are significantly smaller in magnitude than the regular MLEs for all items. The DPLEs are closer to zero than the PLEs. The ridge effects are clearly demonstrated in the items "Place Orientation" and "Time Orientation", in which there are about 16% and 14% reduction of estimates respectively in the DPLE method compared to the PLE method. This is because the two items are highly correlated as mentioned early and the ridge parameter penalized the multicollinearity effect.

In this dataset, finite MLEs were available in the regular logistic regression model given the sample size of 59. However, based on our converged simulation results we expect the MLEs to have larger mean squared errors than either the PLEs or the DPLEs.

Table 3: Parameter estimates and asymptotic standard errors in logistic regression for dementia using MLE, PLE and DPLE.

	MLE		PLE		DPLE(λ=0.01)	
Covariate	Estimate	SE	Estimate	SE	Estimate	SE
Remember Name	-3.63	1.57	-2.05	1.00	-1.97	0.95
Mayor	-2.07	2.28	-0.80	1.39	-0.81	1.31
Civil War	1.31	2.42	0.33	1.57	0.28	1.38
President	-3.95	1.95	-1.99	1.09	-1.91	1.05
Governor	1.57	1.67	0.97	1.04	0.91	1.01
Place Orientation	-1.07	1.05	-0.43	0.70	-0.36	0.63
Time Orientation	-1.03	1.06	-0.69	0.70	-0.59	0.63
Repeat Story	0.85	0.92	0.45	0.56	0.43	0.53
Recall Story	-1.02	0.66	-0.56	0.43	-0.53	0.40
Animals	-0.25	0.28	-0.09	0.16	-0.08	0.14

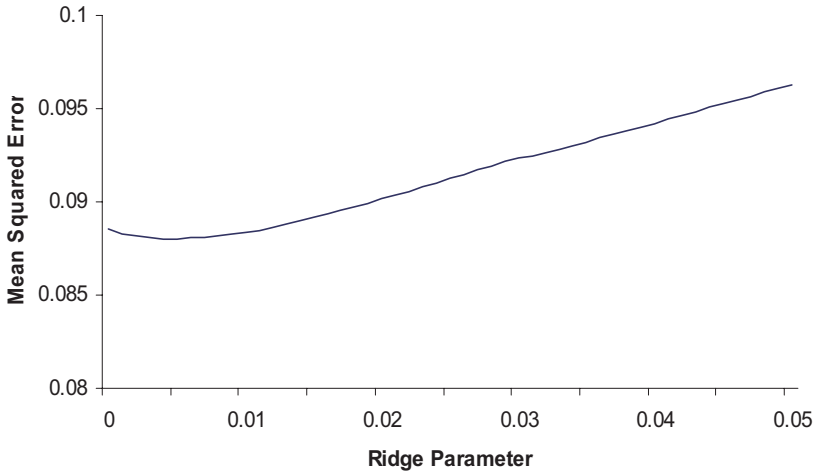

Figure 2: Cross-validated estimates of the mean squared errors as a function of the ridge parameter

5. Discussion

In this paper, we proposed a double penalized maximum likelihood estimator for logistic regression model, providing a solution to the problems of separation and multicollinearity for multivariate item selection in dementia. We demonstrated with simulation results that the double penalized estimator achieves minimum mean squared error at the expense of tolerable amount of bias in small to moderate samples. Most importantly, the double penalized estimation approach yields viable estimates in cases which maximum likelihood estimates do not, therefore providing an attractive alternative to maximum likelihood estimator in logistic regression when a large number of covariates have to be considered. In addition, we acknowledge that the double penalized approach can be applied to any GLIM model though we focused on our discussion in logistic regression model.

Acknowledgement

This research was supported by NIH grants: R01 AG15813, R01 AG09956 and P30 AG10133. We thank Dr.Shelley Bull for providing us with a GAUSS program for comparison with our program.

References

Albert, A. and Anderson, J. A. (1984). On the existence of maximum likelihood estimates in logistic regression models. *Biometrika* **71**, 1-10.

Anderson, J. A. and Richardson, S. C. (1979). Logistic discrimination and bias correction in maximum likelihood estimation. *Technometrics* **21**, 71-78.

Bull, S. B. Mak, C. and Greenwood, C. M. T. (2002). A modified score function estimator for multinomial logistic regression in small samples. *Computational Statistics & Data Analysis* **39**, 57-74.

Cook, R. D. and Weisberg, S. (1982). *Residuals and Influence in regression*. Chapman and Hall.

Copas, J. B. (1988). Binary regression models for contaminated data, with discussuion. *Journal of Royal Statistics Society Series B* **50**, 225-265.

Cordeiro, G. M. and McCullagh, P. (1991). Bias correction in generalized linear models. *Journal of Royal Statistics Society Series B* **53**, 629-643.

Efron, B. (1986). How biased is the apparent error rate of a prediction rule? *Journal of the American Statistiscal Association* **81**, 461-470.

Firth, D. (1992). Bias reduction, the Jeffreys prior and GLIM, In *Advances in GLIM and Statistical Modelling* (Edited by Fahrmeir, L., Francis, B., Gilchrist, R., Tutz, G.). Springer.

Firth, D. (1993). Bias reduction of maximum likelihood estimates.*Biometrika* **80**, 27-38.

Gao, S. and Shen, J. (2007). Asymptotic properties of a double penalized maximum likelihood estimator in logistic regression. *Letters of Probability and Statistics* **77**, 925-930.

Hall, K. S. *et al.* (1996). A cross-cultural community based study of dementias: methods and performance of the survey instrument Indianapolis, U.S.A and Ibadan, Nigeria. *International Journal of Methods in Psychiatric Research* **6**, 129-142.

Heinze, G. and Schemper, M. (2002). A solution to the problem of separation in logistic regression. *Statistics in Medicine* **21**, 2409-2419.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **12**, 55-67.

Hoerl, A. E. and Kennard, R. W. (1988). Ridge Regression, Encyclopedia of Statistical Sciences 9 vols. Plus Supplement, Vol 8, page 129-136, Wiley, New York.

Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proc. R. Soc. A.* **186**, 453-61.

le Cessie, S. and van Houwelingen, J. C. (1992). Ridge estimators in logistic regression. *Applied Statistics-Journal of the Royal Statistical Society Series C* **41**, 191-201.

McLachlanG, J. (1980). A note on bias correction in maximum likelihood estimation with logistic discrimination. *Technometrics* **22**, 621-627.

McCullagh, P. and Nelder, J. A. (1989). *Generalized linear models.* 2nd Edition. Chapman and Hall.

SAS Propriety Software Version 8.2. (2001). SAS institute Inc. Cary, NC.

Schaefer, R.L. Roi, L.D. and Wolfe, R. A. (1984). A ridge logistic regression. *Communications in Statistics - Theory and Methods* **13**, 99-113.

Schaefer, R. L. (1983). Bias correction in maximum likelihood logistic regression. *Statistics in Medicine* **2**, 71-78.

Standardized Occupational Classification Manual. (1980). Washinton, DC: Office of Federal Statistical Policy and Standards.

Received November 16, 2006; accepted June 4, 2007.

Jianzhao Shen
Division of Biostatistics, Department of Medicine
Indiana University School of Medicine
1050 Wishard Blvd. RG4101
Indianapolis, IN 46202-2872, USA
jiashen@iupui.edu

Sujuan Gao
Division of Biostatistics, Department of Medicine
Indiana University School of Medicine
1050 Wishard Blvd. RG4101
Indianapolis, IN 46202-2872, USA
sgao@iupui.edu