DOKUZ EYLÜL UNIVERSITY

GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

PENALIZED LOGISTIC REGRESSION

by

Dinçer GÖKSÜLÜK

July, 2011

İZMİR

PENALIZED LOGISTIC REGRESSION

A Thesis Submitted to the
Graduate School of Natural and Applied Sciences of Dokuz Eylül University
In Partial Fulfillment of the Requirements for the Master of Science in Statistics

by

Dinçer GÖKSÜLÜK

July, 2011

İZMİR

M.SC THESIS EXAMINATION RESULT FORM

We have read the thesis entitled **"PENALIZED LOGISTIC REGRESSION"** completed by **DİNÇER GÖKSÜLÜK** under supervision of **ASSOCIATE PROF. AYLİN ALIN** and we certify that in our opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

ASSOCIATE PROF. AYLİN ALIN

Supervisor

Prof.Dr. Jerna Kurz

(Jury Member)

(Jury Member)

Prof.Dr. Mustafa SABUNCU

Director

Graduate School of Natural and Applied Sciences

ACKNOWLEDGEMENTS

I wish to express my sincere gratitude to my supervisor Associate Prof. Aylin ALIN for her guidance throughout the course of this study.

I also wish to express my gratitude to my family for their supports during my education.

Dinçer GÖKSÜLÜK

PENALIZED LOGISTIC REGRESSION

ABSTRACT

Logistic regression (LR) is frequently used modeling technique for categorical response variables in statistical researches. Binary data are the most common form of categorical response for which the binary outcomes “success” or “failure”, “yes” or “no”. The estimation of regression parameters and classification rate is not accurate when there is multicollinearity among the predictors. In this thesis, we study the penalized logistic regression (PLR) model with quadratic penalization to eliminate the multicollinearity problem and improve the classification rate. We concentrate on several measures for determining the optimum amount of penalization on logistic regression model. We model the real data, coronary heart attack disease data, by both the PLR and LR model and compare their performances.

Keywords: Logistic Regression, Multicollinearity, Newton-Raphson Algorithm, Penalized Logistic Regression, Quadratic Penalization.

CEZALANDIRILMIŞ LOJİSTİK REGRESYON

ÖZ

Lojistik regresyon kategorik verilerin modellemesinde sıklıkla kullanılan bir istatistiksel tekniktir. Kategorik verilerin en yaygın formu "başarılı" veya "başarısız", "evet" veya "hayır" gibi ikili kategorilerin olduğu durumlardır. Regresyon modelini oluşturan değişkenler arasında çoklu doğrusal bağlantı olması durumunda, regresyon modelinin başarı oranı önemli ölçüde düşmektedir. Bu çalışmada, çoklu doğrusal bağlantı sorununu gidermek ve modelin başarı oranını arttırmak için karesel cezalandırılmış lojistik regreson modeli kullanılmıştır. En uygun cezalandırma miktarını belirlemek için çeşitli ölçüler kullanılmıştır. Bu iki yöntem gerçek veri setine (koroner kalp krizi verileri) uygulanmış ve performansları bakımından karşılaştırmaları yapılmıştır.

Anahtar Kelimeler: Lojistik Regresyon, Çoklu Doğrusal Bağlantı, Newton-Raphson Algoritması, Cezalandırılmış Lojistik Regresyon, Quadratic Cezalandırma.

CONTENTS

Page

CHAPTER ONE

INTRODUCTION

Logistic regression (LR) is a special form of General Linear Models (GLMs). Although it's mostly used with binary responses, we can use LR models with multi-category responses (more than two categories). The categories might be either in ordinal scale or nominal scale. The choice of method depends on the scale of the response variable. We can select the best criteria with respect to properties of response variable and covariates. For all cases, the LR method models the success probabilities of dependent variable.

The regression parameters are estimated by using several estimation methods. Maximum likelihood estimation method (MLE) is frequently used for GLMs. In practice, the data might be stored in different forms. The sample observations might be grouped or ungrouped. When the data is grouped, we use contingency tables to predict the cell probabilities. Thus, the response variable is distributed as binomial. The logistic regression parameters are estimated by using the binomial likelihood equations. If the data is ungrouped, the likelihood function of Bernoulli distribution for response is used to obtain the regression coefficients.

Logistic regression model is not appropriate to fit the data when there is multicollinearity among the explanatory variables or when the number of explanatory variables is relatively large. In both of these scenarios, the estimated parameters become unstable. The regression model might be very poor for predicting new observations. Thus, we should carefully consider the high dimension and multicollinearity problems before estimating the regression coefficients. In literature, there are many modifications to improve the parameter estimates.

Penalized logistic regression (PLR) is a method which is based on the idea that penalizing the unstable regression coefficients to obtain robust regression coefficients to multicollinearity problem. This method introduces a penalty parameter to the likelihood function. We may estimate regression coefficients by maximizing the penalized likelihood function at the optimum level of penalty parameter. In literature, there are several penalization methods. We propose the quadratic penalization (ridge penalization) in this thesis to overcome the multicollinearity problem. This method estimates the robust regression parameters against multicollinearity. However, the estimated coefficients are not unbiased anymore.

Cessie and Houwelingen (1990) propose ridge estimators (quadratic penalization) in logistic regression to improve the parameter estimates and to overcome both the multicollinearity and high dimension problem. This paper provides useful measures based on cross-validation to define the amount of penalization (ridge parameter). Zhu and Hastie (2004) study two different methods PLR based on quadratic penalization and Support Vector Machine (SVM) for cancer classification from microarray data set. Three different data sets are used to compare the classification rate of PLR and SVM methods. Aguilera *et al.* (2006) study another alternative method Principal Component Analysis (PCA) for estimating the logistic regression model with high dimensional and multicollinear data. A simulated data set is also used to examine the PCA method. Park and Hastie (2007) propose quadratic penalization for detecting gene interactions. The advantages of quadratic penalization are also provided in this study. They studied both simulated and real data sets with PLR method. Shen and Tan (2005) study the combination of two dimension reduction method, PLS and singular value decomposition (SVD), with the penalized logistic regression to provide a powerful classification for cancer diagnosis using the microarray data. These methods are compared on the data set given by Golub *et al.* (1999). We can give many examples for logistic regression in medical applications. In this thesis, we study with clinical data to classify the binary response by PLR.

This thesis contains three main chapters. In chapter two, we give detailed explanation on binary logistic regression method where the response variable has only two categories. We focused on how to estimate the regression coefficients with MLE method. The general information and the inferences for logistic regression is given in this chapter. In chapter three, we propose PLR with quadratic penalization to overcome multicollinearity problem. We give general information about Newton-Raphson algorithm to obtain parameter estimates. The advantages and disadvantages of PLR method is also considered in chapter three. Chapter four includes the numerical study. We compare binary LR and PLR on the same data set. A real data set, coronary heart attack data, is used to compare the classification rate of LR and PLR. Finally, we conclude our study with chapter five. The MATLAB codes and data set are given in Appendix.

CHAPTER TWO

LOGISTIC REGRESSION

Logistic regression, as an extension of standard regression analysis, is widely used modeling technique for categorical response variables in statistical researches. Binary data are the most common form of categorical responses, for which the possible outcomes “success” or “failure”, “yes” or “no” (Agresti, 2007, p.99). It is used in social, educational and medical sciences. In educational researches, classifying the students whether learning disabled or succeed in college is a use of binary logistic regression. Similarly, in financial researches, defining the credit risk which is the probability that a customer is credit worthy or not is another use of binary logistic regression. Recently, logistic regression has become one of the most popular tools in medical applications. Although it’s mostly called with medical applications, logistic regression is useful modeling technique in business and marketing. For instance, a company may wish to know that to open a new office at a new place or not. The economical position of the public in this area, educational status of people, the potential number of the customers that live nearby and past buying behaviors of people might be independent variables for modeling the marketing strategies.

Another important area of logistic regression applications is genetics. In microarray applications, it is the most preferable statistical modeling. Many recent articles that concern with genetic variations published for classifying cancer using the DNA information.

In this chapter we study the binary logistic regression more closely. Model building and interpreting the unknown parameters will be considered.

2.1 Simple Logistic Regression

Let Y be a binary response variable with categories "success" and "failure". Before we build the logistic regression model, the categories of response variable are recoded as 0 for failure and 1 for success. Let *X* be the explanatory variable. The ordinary least squares (OLS) model for response is defined as:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \qquad i = 1,2,\ldots,n \tag{2.1}$$

where X_i is explanatory variable and the value for *i*-th case in (2.1), that is Y_i , is Bernoulli distributed binary response with categories 0 and 1. Thus, the expected mean of response can be written as:

$$E[Y_i] = \beta_0 + \beta_1 X_i \qquad i = 1,2,\ldots,n \tag{2.2}$$

where $E[\varepsilon_i] = 0$. Since Y_i is Bernoulli random variable, its probability distribution can be written as follows:

$$f(Y;\pi) = \begin{cases} \pi & if \quad Y = 1, \\ 1-\pi & if \quad Y = 0, \\ 0 & otherwise. \end{cases} \tag{2.3}$$

Thus, π_i is the probability that $Y_i = 1$ and $(1-\pi_i)$ is the probability that $Y_i = 0$. The expected value of Bernoulli random response *Y* is:

$$E[Y_i] = \beta_0 + \beta_1 X_i = 1(\pi_i) + 0(1-\pi_i) = \pi_i \quad (2.4)$$

where the mean response is the probability of "success" when the response is binary categorical variable. Modeling the probability requires to be very careful. Since the probability must have continuous values within 0 and 1, our regression model must generate mean responses within the interval $0 \le E[Y_i] \le 1$. However, the OLS method generates mean responses beyond this interval, which is the greater value than upper limit 1 or smaller value than lower limit 0 for different X values. Because the mean response is bounded, OLS method is not appropriate for binary responses. The mean response must have values within the interval [0,1] while explanatory variables have no constraints. Thus, the relation between the variables is not linear. We can graphically define the relation as follow:

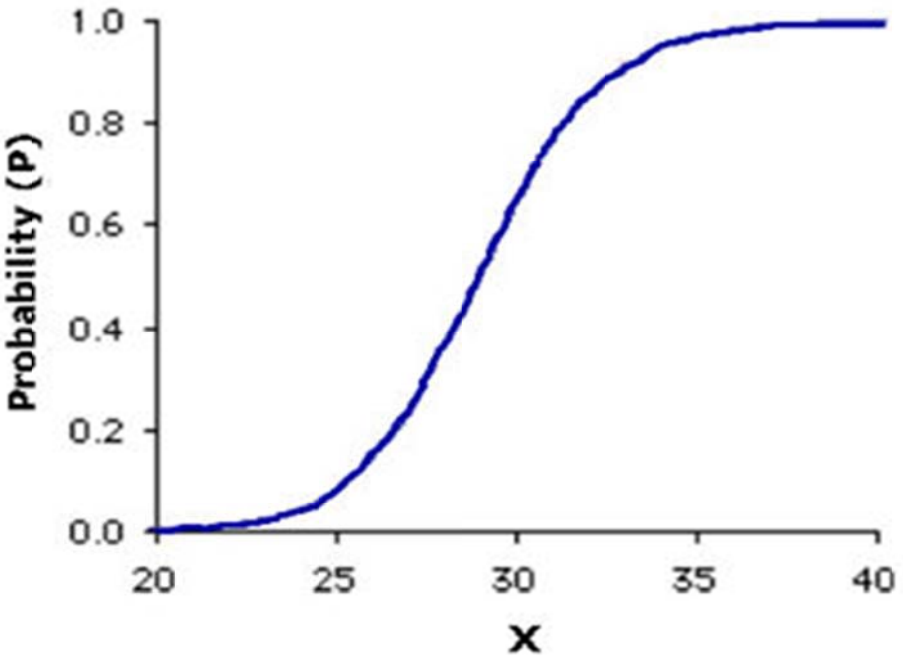


Figure 2.1 Logistic regression curve. The relation between response and predictors is S-shaped.

As we can see from Figure 2.1, there is an S-shaped relation between explanatory variable and the mean response. Logit transformation is the best form for modeling this relation. The logistic regression model can be expressed as:

$$\pi_i = E[Y_i] = \frac{\exp(\beta_0 + \beta_1 X_i)}{1 - \exp(\beta_0 + \beta_1 X_i)} \tag{2.5}$$

where β_0 denotes the intercept.

While the success probability is near 0 or 1, per unit change in explanatory variable has smaller effect in the magnitude of mean response (probability). It's generally harder to interpret the logistic curve than linear equations. However, we can make linear approximation to logistic regression curve by using *logit* transformation which is expressed by Hosmer and Lemeshov. Logit transformation generates new mean responses within the interval $[-\infty, \infty]$ which are not a success probability anymore. Equivalently, the linear logit model or *log odds* is:

$$\text{logit}[\pi(x)] = \log \frac{\pi(x)}{1 - \pi(x)} = \beta_0 + \beta_1 x \tag{2.6}$$

where $\pi(x) = P(Y = 1 \mid X = x)) = 1 - P(Y = 0 \mid X = x))$. This equates the logit link function to the linear predictor (Agresti, 2002, p.166).

2.2 Multiple Logistic Regression

As a general form of logistic regression models, the simple logistic regression can be easily extended to more than one explanatory variable which is called multiple logistic regression model. In fact, several predictors might be required to obtain

better interpretations and fits. The multiple logistic regression model with *k* predictors is given with equation (2.7).

$$\text{logit} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k \tag{2.7}$$

The interpretations for regression coefficients, odds and logit are similar to the simple logistic regression. The parameter $\beta_i \quad (i = 0,1,...,k)$ indicates the effect of X_i on the logit or log odds at fixed levels of other predictors. In this model, estimated probabilities at point *x* denoted as:

$$\pi(x) = \frac{\exp(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik})}{1 + \exp(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik})}. \tag{2.8}$$

The explanatory variables can be continuous or categorical. If a predictor is categorical, it is included in the model as *dummy* variable. When all variables are categorical, the data can be grouped and displayed in a multiway contingency table. In this situation, the multiple logistic regression model refers to the *log-linear* models. Suppose a categorical variable with *c* categories, the multiple regression model contains (*c*-1) dummy variables since one category is selected as base category. In statistical packages, if the categorical variable has a natural ordering, first category or last category is selected as baseline. Researchers may decide the selection of baseline category.

2.2.1 Matrix Approach to Logistic Regression

We have a sample with size *n* and *k* predictors. There is an *n* paired set $(x_i, y_i) \quad i = 0,1,2,\ldots,n$. The response is assumed to be binary variable. To simplify the calculations, we will use matrix approach to the multiple logistic regression model. Refer to equation (2.7), we define the vectors and matrices as follows:

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}_{(k+1)\times 1} \qquad \mathbf{X} = \begin{bmatrix} 1 & x_{12} & x_{13} & \cdots & x_{1k} \\ 1 & x_{22} & x_{23} & \cdots & x_{2k} \\ 1 & x_{32} & x_{33} & \cdots & x_{3k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n2} & x_{n3} & \cdots & x_{nk} \end{bmatrix}_{n\times(k+1)} \qquad \boldsymbol{\eta} = \begin{bmatrix} \eta_1 \\ \eta_2 \\ \vdots \\ \eta_n \end{bmatrix}_{n\times 1} \tag{2.9}$$

Here, $\boldsymbol{\beta}$ denotes the coefficients vector and $\mathbf{X}$ denotes the matrix of observations. $\boldsymbol{\eta}$ is the column vector of logits for *i*th observation. Notice that first column of the matrix $\mathbf{X}$ has a vector of 1 which indicates the regression constant β_0. Now we can express our model with matrix notations as:

$$\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{2.10a}$$

$$\begin{bmatrix} \eta_1 \\ \eta_2 \\ \vdots \\ \eta_n \end{bmatrix}_{n\times 1} = \begin{bmatrix} 1 & x_{12} & x_{13} & \cdots & x_{1k} \\ 1 & x_{22} & x_{23} & \cdots & x_{2k} \\ 1 & x_{32} & x_{33} & \cdots & x_{3k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n2} & x_{n3} & \cdots & x_{nk} \end{bmatrix}_{n\times(k+1)} \cdot \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}_{(k+1)\times 1} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}_{n\times 1} \tag{2.10b}$$

where $\boldsymbol{\varepsilon}$ denotes the column vector of error terms.

2.3 Interpreting Model Parameters

Interpreting model parameters in logistic regression should be carefully considered. Because the logistic model has complex and non-linear form, interpreting the model parameters might be more difficult comparing with linear models. Although linear approximation to logistic regression curve makes it easier, interpretation still requires more attention. From linear logit model, the sign of β determines the direction of the logistic curve. For $\beta = 0$, the response variable is

independent from predictors. The effect of β on the regression curve is graphically shown as follows:

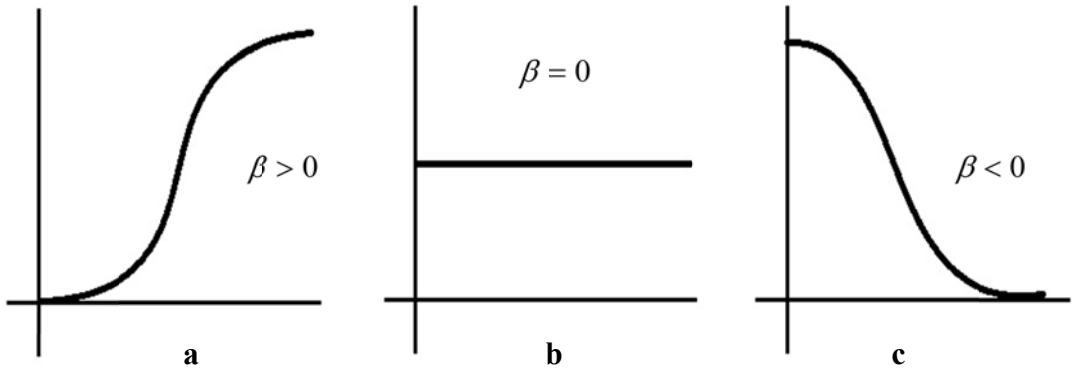


Figure 2.2 The effect of regression parameters to the logistic regression curve. a) Positive relation, b) No relation, c) Negative Relation.

The per unit change in explanatory variables has increasing or decreasing effect in the logit or base logarithm of odds. Similarly, it has same effect on the success probability. However, these effects can't be measured directly with the magnitude of unknown parameter β. From equation (2.6), 1-unit change in predictor variable indicates that the logit or log odds increases by β. However, measuring the change in logit may not be familiar for most scientists. Some alternative interpretations can be considered. Exponentiating both sides of (2.6) we get the following equation:

$$\text{odds} = \frac{\pi(x)}{1-\pi(x)} = \exp(\beta_0 + \beta_1 x) = e^{\beta_0}(e^{\beta_1})^x \tag{2.11}$$

We can see that the odds are expressed as exponential form of predictors. Since e^{β_0} is constant, the odds increase multiplicatively by e^{β_1} for every unit change in predictor variables which is called *odds ratio*. Odds ratio is a division of two odds at the points x and $x+1$, respectively. Odds ratio indicates the magnitude of increase or

decrease in odds for every unit change in predictors. This can be obtained mathematically as follows:

$$\text{OR} = \frac{odds(x+1)}{odds(x)} = \frac{\dfrac{\pi(x+1)}{1-\pi(x+1)}}{\dfrac{\pi(x)}{1-\pi(x)}} = \frac{\exp(\alpha+\beta_1(x+1))}{\exp(\alpha+\beta_1 x)} = \frac{e^{\alpha}(e^{\beta_1})^{x+1}}{e^{\alpha}(e^{\beta_1})^{x}}$$

$$\text{OR} = \frac{e^{\alpha}(e^{\beta_1})^{x}e^{\beta_1}}{e^{\alpha}(e^{\beta_1})^{x}} = e^{\beta_1}. \tag{2.12}$$

Another alternative measure for a good interpretation is to obtain the *slope* of regression curve at a given point of *x*. This provides us to calculate the rate of probability change at that point. Increasing the magnitude of predictor from *x* to *x*+1 changes the success probability with a rate of $\beta\pi(x)[1-\pi(x)]$. Since the rate of a function is the derivation of corresponding function, the rate of change in $\pi(x)$ curve can be mathematically calculated by taking derivative $\partial\pi(x)/\partial x$ using the equation (2.5). The slope decreases while success probability approaches 0 or 1. The maximum slope occurs at the *x* point for which the success probability equals to 1/2. The per unit change at that point will change the success probability as $\beta(1/2)(1/2) = 0.25\beta$. This *x* level is called *median effective level* and denoted EL_{50}. The median effective level can be approximated as $-\alpha/\beta$.

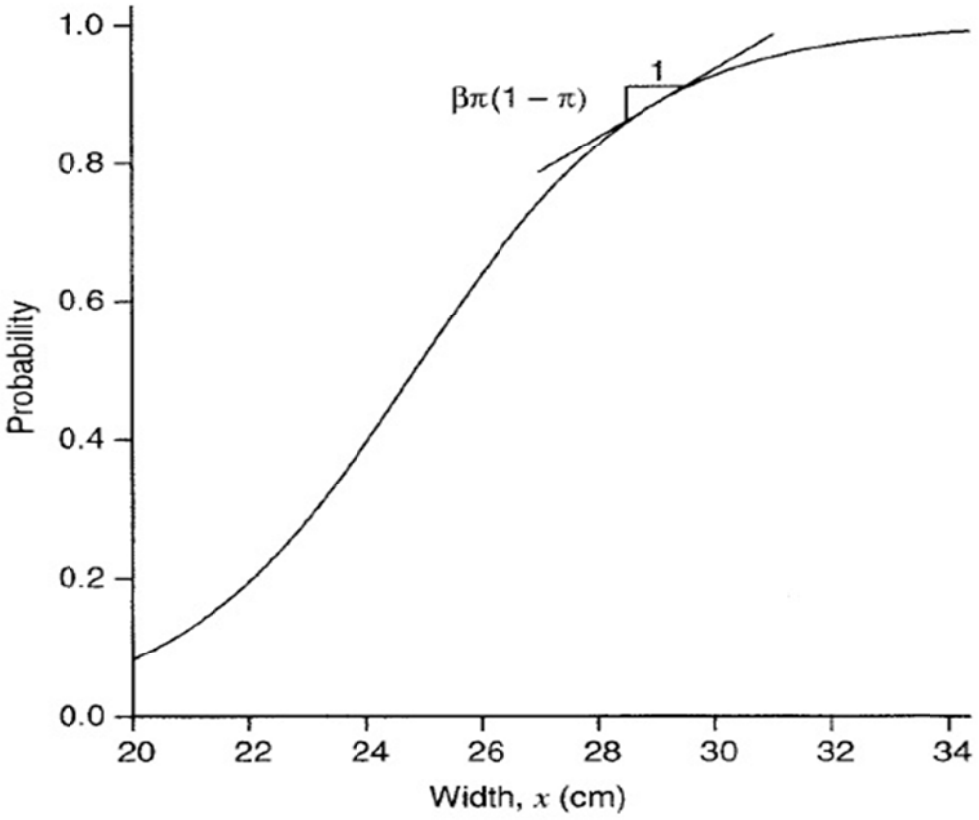


Figure 2.3 Linear approximation to logistic regression curve. The slope of the regression curve at a given level of x can be calculated with tangent line which is drawn to the regression curve. (Agresti, 2002, p.167)

2.4 Inference for Logistic Regression

The inference procedures in logistic regression is similar to the simple linear regression -inferences about regression coefficients, intervals for mean response, prediction of new observations and its interval estimations. Since logistic regression curve is not linear, the inferences for regression coefficients might be more complex.

2.4.1 Inference for Effects

Let our model be the logit model with several predictors. Suppose our sample is large enough. For large samples, maximum likelihood estimators for predictors are approximately normally distributed. The estimated variances for regression parameters can be obtained from second-order partial derivatives of the log-

likelihood function. Under the large sample consideration, Wald confidence interval for the parameter β in the logit model is:

$$\hat{\beta} \pm z_{\alpha/2}(SE). \tag{2.13}$$

If the sample size is small or fitted probabilities are located near 0 and 1 even the sample is large enough, Wald confidence interval is not adequate. The likelihood ratio based confidence interval is preferable instead of Wald interval. If confidence interval for β is estimated, it can be updated into confidence interval for logits, odds ratio, slope and success probability. The confidence interval for odds ratio is equal to $[\exp(\hat{\beta} - z_{\alpha/2}(SE)), \exp(\hat{\beta} + z_{\alpha/2}(SE))]$. Similarly, the confidence interval for slope at $\pi(x) = 0.50$ is $0.25[\hat{\beta} \pm z_{\alpha/2}(SE)]$.

Although we don't need to make inference for regression constant α, it is used while making inference for event probabilities and logits. Before we make an inference for logit, we need to calculate the estimated variance of logits. From our logit model, the estimated variance is:

$$\text{var}\left[\text{logit}\left[\pi(\text{x})\right]\right] = \text{var}\left(\hat{\alpha} + \hat{\beta}x\right) = \text{var}(\hat{\alpha}) + x^2\,\text{var}(\hat{\beta}) + 2x\,\text{cov}(\hat{\alpha}, \hat{\beta}) \tag{2.14}$$

The SE of logit is the square root of equation (2.14). Hence, the estimated confidence interval of logit at point *x* is:

$$\text{logit}\left[\pi(\text{x})\right] \pm z_{\alpha/2}(SE) \tag{2.15}$$

As we mentioned earlier, most scientists are not familiar with logits. Researchers may wish to know about event probabilities. Once logit is obtained, we can upgrade confidence interval for probabilities by transforming the estimated confidence interval of logits. Let the confidence interval of logit at *x* is $[\mathrm{LL}, \mathrm{LU}]$, LU for upper limit and LL for lower limit, the corresponding confidence interval for $\pi(\mathrm{x})$ is:

$$\left[\frac{\exp(\mathrm{LL})}{1+\exp(\mathrm{LL})}, \frac{\exp(\mathrm{LU})}{1+\exp(\mathrm{LU})}\right] \tag{2.16}$$

Determining confidence interval for probabilities has a minor exception. One could obtain this interval by using the sample proportion instead of fitted model. If the observations are repeated at *x*, the sample proportion at this point might be used for estimation of probabilities. Because the repeated observations are distributed as Binomial, SE of $\pi(\mathrm{x})$ is calculated from Binomial distribution as $\sqrt{\hat{\pi}(\mathrm{x})(1-\hat{\pi}(\mathrm{x}))/\mathrm{n}}$. In most cases, these two intervals will be different from each other. If the repeated samples are not enough, using the full model will produce more consistent confidence intervals.

2.4.2 Significance Testing of Parameters

For logistic regression model, selecting the redundant parameters is cruel. Before performing the significance tests, we should understand the data set clearly. For large samples, Wald statistics and likelihood based statistics can be selected. Wald statistics are less complex comparing to the Likelihood based statistics. For null hypothesis $\mathrm{H}_0 : \beta = 0$, the Wald statistic indicates that the corresponding predictor should be removed from logistic regression model or not. For a single predictor, the following statistic has a standard normal distribution under the null hypothesis.

$$z = \hat{\beta} / SE \tag{2.17}$$

One refers z to the standard normal table to get a one-sided or two-sided P-value (Agresti, 2007). For the single predictor and two-sided $H_a : \beta \neq 0$, $z^2 = (\hat{\beta}/SE)^2$ has a chi squared null distribution with one degrees of freedom. This type of statistic using the non-null standard error is called a Wald statistic (Agresti, 2007, p:11). Although the Wald test is adequate for large samples, the likelihood based test statistics are more powerful for large samples. The likelihood-ratio test is based on the magnitudes of likelihood function. This statistic compares the maximum value of the log-likelihood function under the null hypothesis to the value of the log-likelihood under the constraints of alternative hypothesis. The likelihood-ratio test and Wald test give similar results for large samples. However, likelihood ratio test is more preferable than Wald because it uses more information.

2.5 Model Building and Variable Selection

Most of the regression methods include more than one predictor variable. When there are several variables in the model, the strategies of selecting important predictors are crucial. In regression methods there are several ways of excluding the redundant predictors. Stepwise method, Backward elimination, Forward Selection and Best Subset Regression Method might be reviewed as commonly used variable selection methods. In this section, we will focus on the methods for testing several variables at once. Equivalently, we will decide whether a set of variables should be retained or removed from the model. Introducing the variable selection algorithms is out of the scope of this research.

As we previously mentioned, Wald test is used for significance of single coefficient. When a set of variables are being tested, alternative test statistics should be preferred. The likelihood ratio (LR) test is one of the common test statistics when testing a set of predictors. This test is based on a statistic called *deviance*. The deviance of a model is an extension of the magnitude of likelihood function.

Suppose a large sample with *p* unknowns and *n* observations. The logistic regression model can be fitted perfectly if we estimate a parameter for each individual which leads us to estimate *n* different parameter. This model, which is called *saturated model*, includes all the information and provides perfect fit. Equivalently, the error term for this model is zero. Thus, the likelihood for saturated model is 1 and so that the log-likelihood is equal to 0:

$$\log_e L_S = \sum_{i=1}^{n} [Y_i \log_e(Y_i) + (1 - Y_i) \log_e(1 - Y_i)] = 0 \tag{2.18}$$

This log-likelihood value for saturated model is compared with the log-likelihood value of fitted model.

$$\log_e L_F(\beta_0, \beta_1, \ldots, \beta_{p-1}) = \sum_{i=1}^{n} Y_i(\mathbf{X}\boldsymbol{\beta}) - \sum_{i=1}^{n} \log_e[1 + \exp(\mathbf{X}\boldsymbol{\beta})] \tag{2.19}$$

The log-likelihood for fitted model can never be larger than the saturated model because it has fewer parameters. The difference between these likelihood values is equal to its *deviance* which is used as a goodness of fit criterion. Thus, the model deviance can be written as follow:

$$DEV(X_0, X_1, \ldots, X_{p-1}) = 2 \log_e L_S - 2 \log_e L_F(\beta_0, \beta_1, \ldots, \beta_{p-1}) \tag{2.20}$$

Since, log-likelihood for saturated model is 0, the model deviance can be expressed as:

$$
\begin{aligned}
DEV(X_0, X_1, \ldots, X_{p-1}) &= -2\log_e L_F(\beta_0, \beta_1, \ldots, \beta_{p-1}) \\
&= -2\sum_{i=1}^{n}[Y_i \log_e(\hat{\pi}_i) + (1-Y_i)\log_e(1-\hat{\pi}_i)] \\
&= -2\sum_{i=1}^{n} Y_i(\mathbf{X\beta}) - \sum_{i=1}^{n} \log_e[1+\exp(\mathbf{X\beta})]
\end{aligned}
\tag{2.21}
$$

where $\hat{\pi}_i$ is the fitted probability for *i*th observation. This statistic has chi-square distribution with degrees of freedom $(n-1)-(p-1)=n-p$. Equivalently, the deviance is compared with the value $\chi^2(1-\alpha; n-p)$ from chi-square table. The larger the model deviance, the poorer is the fit (Neter and Kutner, 1996, p.587).

The deviance methodology can be used for testing several unknown parameters. Suppose we have *p* unknown parameters, one may wish to conduct a test such as:

$$
\begin{aligned}
&H_0 : \beta_q = \beta_{q+1} = \cdots = \beta_{p-1} = 0 \\
&H_a : not \quad all \quad \beta's \quad in \quad H_0 \quad equal \quad zero
\end{aligned}
$$

First, we need to define the fitted model which includes all predictors that is *full model*:

$$
\text{logit} = \beta_0 + \beta_1 X_1 + \cdots + \beta_{p-1} X_{p-1} \tag{2.22}
$$

and the fitted model which includes the unknowns except those given with the null hypothesis; *reduced model:*

$$\text{logit} = \beta_0 + \beta_1 X_1 + \cdots + \beta_{q-1} X_{q-1} \tag{2.23}$$

Now we can denote the deviance of the full model $DEV(X_0, X_1, \ldots, X_{p-1})$ and the reduced model $DEV(X_0, X_1, \ldots, X_{q-1})$. The difference between these deviances determines the significance of several predictors at once. If the difference is large, the predictors given with the null hypothesis should be retained in the model because they improve the fit substantially.

$$DEV(X_q, \ldots, X_{p-1} \mid X_0, \ldots, X_{q-1}) = DEV(X_0, \ldots, X_{q-1}) - DEV(X_0, \ldots, X_{p-1}) \tag{2.24}$$

This difference between two deviances is called *partial* deviance. Notice that the deviances of the full model and reduced model have initially been compared with the saturated model. The degrees of freedom for reduced model is $(n-1)-(q-1) = n-q$ and the degrees of freedom for the full model is $(n-1)-(p-1) = n-p$. Thus, the partial deviance has chi-square distribution with degrees of freedom $(n-q)-(n-p) = p-q$. Notice that the analogy of deviance to the extra sum of squares for linear regression models. Deviance statistic is used either testing single predictors or a set of predictors. Hence, we are able to compare several models with each other and decide the best model among the available subsets.

Suppose we wish to test a single predictor $H_0 : \beta_1 = 0$, that is, the predictor X_1 should be dropped from the model or not, the partial deviance is calculated as follows:

$$DEV(X_1 \mid X_0, X_{2,}, \ldots, X_{p-1}) = DEV(X_0, X_2, \ldots, X_{p-1}) - DEV(X_0, X_1, \ldots, X_{p-1}) \quad (2.25)$$

where the partial deviance has chi-square distribution with $df = 1$ that is $[(n-1)-(p-2)]-[(n-1)-(p-1)] = 1$. Large values of the partial deviance lead to conclusion of retaining predictor in the model.

Model deviance and partial deviances are important for model building and variable selection procedure. We can easily decide which model is best and which variables should be included or dropped from the model with the help of deviance statistics.

2.6 Fitting Logistic Regression Models

In statistics several model fitting techniques are developed. However, we should be careful what to use for fitting the model. The data properties and model assumptions directly affect the method selection process. Thus, we should understand the data correctly before fitting the model.

In linear regression model, the least squares method is used to obtain the parameter estimates. It is based on minimizing the sum of squared errors of the predicted values. This method is not appropriate for logistic regression models since the response is binary. Similarly, the error term is assumed to be randomly normally distributed. However, in logistic regression the error term is not random because it equals to $1-\pi(x_i)$ when $Y = 1$ and $-\pi(x_i)$ when $Y = 0$. The commonly used estimating method for logistic regression models are:

1. The Maximum Likelihood Method (MLE).
2. Iteratively Reweighted Least Squares Method.
3. The Minimum Logit Chi-Square Method.

In this paper, the method of maximum likelihood (MLE) is used to estimate the unknown regression parameters.

2.6.1 Maximum Likelihood Method (MLE)

The most commonly used estimation method for logistic regression model is the *method of maximum likelihood*. The likelihood function of response is equal to its joint probability distribution.

$$\begin{aligned} g(Y_1, Y_2, \ldots, Y_n) &= \prod_{i=1}^{n} f_i(Y_i) \\ &= \prod_{i=1}^{n} \pi_i^{Y_i} (1-\pi_i)^{1-Y_i} \end{aligned} \tag{2.26}$$

where Y_i is Bernoulli random variable. The log-likelihood is equal to:

$$\begin{aligned} \log\big(g(Y_1, Y_2, \ldots, Y_n)\big) &= \log\left(\prod_{i=1}^{n} \pi_i^{Y_i} (1-\pi_i)^{1-Y_i} \right) \\ &= \sum_{i=1}^{n} Y_i \log(\pi_i) + \sum_{i=1}^{n} (1-Y_i) \log(1-\pi_i) \\ &= \sum_{i=1}^{n} Y_i \log\left(\frac{\pi_i}{1-\pi_i} \right) + \sum_{i=1}^{n} \log(1-\pi_i) \end{aligned} \tag{2.27}$$

Substituting the expressions π_i and $(1-\pi_i)$ into log-likelihood equation, we may write;

$$\begin{aligned} \ln L(\beta) = &\sum_{i=1}^{n} Y_i \left(\beta_0 + \beta_1 X_{i1} + \cdots + \beta_{q-1} X_{i(q-1)} \right) \\ &- \sum_{i=1}^{n} \log\left[1 + \exp\left(\beta_0 + \beta_1 X_{i1} + \cdots + \beta_{q-1} X_{i(q-1)} \right) \right] \end{aligned} \tag{2.28}$$

Differentiating equation (2.28) with respect to β_0 and β_k's we obtain the following equations.

$$\frac{\partial \ln L(\beta)}{\partial \beta_0} = \sum_{i=1}^{n} Y_i - \sum_{i=1}^{n}\left(\frac{\exp\left(\beta_0 + \beta_1 X_{i1} + \cdots + \beta_{q-1} X_{i(q-1)}\right)}{1+\exp\left(\beta_0 + \beta_1 X_{i1} + \cdots + \beta_{q-1} X_{i(q-1)}\right)}\right) \tag{2.29a}$$

$$\frac{\partial \ln L(\beta)}{\partial \beta_a} = \sum_{i=1}^{n} Y_i X_{ia} - \sum_{i=1}^{n}\left(\frac{X_{ia}\exp\left(\beta_0 + \beta_1 X_{i1} + \cdots + \beta_{q-1} X_{i(q-1)}\right)}{1+\exp\left(\beta_0 + \beta_1 X_{i1} + \cdots + \beta_{q-1} X_{i(q-1)}\right)}\right) \tag{2.29b}$$

Setting the likelihood equations (2.29a) and (2.29b) equal to zero, we get the following *nonlinear* functions.

$$\sum_{i=1}^{n} Y_i - \sum_{i=1}^{n}\left(\frac{\exp\left(\beta_0 + \beta_1 X_{i1} + \cdots + \beta_{q-1} X_{i(q-1)}\right)}{1+\exp\left(\beta_0 + \beta_1 X_{i1} + \cdots + \beta_{q-1} X_{i(q-1)}\right)}\right) = \sum_{i=1}^{n}(Y_i - \pi_i) = 0 \tag{2.30a}$$

$$\sum_{i=1}^{n} Y_i X_{ia} - \sum_{i=1}^{n}\left(\frac{X_{ia}\exp\left(\beta_0 + \beta_1 X_{i1} + \cdots + \beta_{q-1} X_{i(q-1)}\right)}{1+\exp\left(\beta_0 + \beta_1 X_{i1} + \cdots + \beta_{q-1} X_{i(q-1)}\right)}\right) = \sum_{i=1}^{n} X_{ia}\left(Y_i - \pi_i\right) = 0. \tag{2.30b}$$

Equation (2.30a) and (2.30b) requires iterative solution. We use Newton-Raphson algorithm to estimate the regression parameters. NR algorithm based on solving the second order Taylor series expression of the likelihood function. Approximating the nonlinear function with second order Taylor series generates linear form of the corresponding function.

Let f be continuous on a closed interval $[a,b]$ and its derivatives $f', f'',..., f^{(q-1)}$ are exist. Let the qth-order derivative $f^{(q)}(x)$ exists for every $x \in (a,b)$. Then there exist $c \in (a,b)$ such that (Fleming, W., 1977, p.386)

$$f(b) - f(a) = f'(a)(b-a) + \frac{f''(a)}{2!}(b-a)^2 + \cdots + \frac{f^{(q-1)}(a)}{(q-1)!}(b-a)^{q-1} + R_q \tag{2.31}$$

where the remainder equals

$$R_q = \frac{f^{(q)}(c)}{q!}(b-a)^q . \tag{2.32}$$

The remainder term is excluded because it rapidly approaches to zero while $q \to \infty$. Let $l(\beta)$ be the log-likelihood and h is a point in real space. The *second-order* Taylor series approximation of $l(\beta)$ on the interval $\left[\beta^t, \beta^t + h\right]$ is expressed as

$$l(\beta^t + h) - l(\beta^t) = \frac{\partial l(\beta^t)}{\partial \beta}(\beta^t + h - \beta^t) + \frac{\partial^2 l(\beta^t)}{\partial \beta^2}(\beta^t + h - \beta^t)^2 \frac{1}{2}. \tag{2.33}$$

The expression $(\beta^t + h)$ denotes the parameter estimation at step t+1 which is denoted by β^{t+1}. For each step the estimation of β^t is known. Thus, we should estimate the value of h to get the parameter estimation of β^{t+1}. The equation (2.33) reduces to

$$l(\beta^t + h) = l(\beta^t) + \frac{\partial l(\beta^t)}{\partial \beta} h + \frac{\partial^2 l(\beta^t)}{\partial \beta^2} h^2 \frac{1}{2}. \tag{2.34}$$

Here, we obtain the parameter estimations by maximizing the function $l(\beta^t + h)$. The maximum of this function is equal to the point $\beta^t + \hat{h}$ that equates the first derivation of $l(\beta^t + h)$ to zero. Hence, we need to obtain maximum likelihood estimate for h. Differentiating $l(\beta^t + h)$ with respect to h we get,

$$\begin{aligned} \frac{\partial l(\beta^t + h)}{\partial h} &= \frac{\partial l(\beta^t)}{\partial \beta} + \frac{\partial^2 l(\beta^t)}{\partial \beta^2} h \\ 0 &= \frac{\partial l(\beta^t)}{\partial \beta} + \frac{\partial^2 l(\beta^t)}{\partial \beta^2} \hat{h} \\ \hat{h} &= -\frac{\dfrac{\partial l(\beta^t)}{\partial \beta}}{\dfrac{\partial^2 l(\beta^t)}{\partial \beta^2}}. \end{aligned} \tag{2.35}$$

Equivalently, the approximated likelihood function $l(\beta^t + h)$ gives its maximum at point $\beta^t + \hat{h}$ that is the parameter estimation at step t+1. We can express this point as follow.

$$\beta^{t+1} = \beta^t + \hat{h} = \beta^t - \frac{\partial l(\beta^t) \big/ \partial \beta}{\partial^2 l(\beta^t) \big/ \partial \beta^2}. \tag{2.36}$$

Since equation (2.36) is solved iteratively, the final estimation is achieved when the function converges. As seen from (2.36), NR method entails determining the

second order derivatives. The second order derivative of log-likelihood with respect to β_0 is:

$$\begin{aligned}\frac{\partial}{\partial\beta_0}\left[\sum_{i=1}^{n}\left(Y_i-\pi_i\right)\right] &= \frac{\partial}{\partial\beta_0}\left[-\sum_{i=1}^{n}\left(\pi_i\right)\right] \\ &= \frac{\partial}{\partial\beta_0}\left[-\sum_{i=1}^{n}\left(\frac{\exp\left(\beta_0+\beta_1X_1+\cdots+\beta_{q-1}X_{q-1}\right)}{1+\exp\left(\beta_0+\beta_1X_1+\cdots+\beta_{q-1}X_{q-1}\right)}\right)\right] \\ &= -\sum_{i=1}^{n}\pi_i\left(1-\pi_i\right).\end{aligned} \tag{2.37a}$$

Similarly, the second order derivative with respect to β_k's ;

$$\begin{aligned}\frac{\partial}{\partial\beta_b}\left[\sum_{i=1}^{n}X_{ia}\left(Y_i-\pi_i\right)\right] &= \frac{\partial}{\partial\beta_b}\left[-\sum_{i=1}^{n}X_{ia}\left(\pi_i\right)\right] \\ &= \frac{\partial}{\partial\beta_b}\left[-\sum_{i=1}^{n}\left(X_{ia}\frac{\exp\left(\beta_0+\beta_1X_{i1}+\cdots+\beta_{q-1}X_{i(q-1)}\right)}{1+\exp\left(\beta_0+\beta_1X_{i1}+\cdots+\beta_{q-1}X_{i(q-1)}\right)}\right)\right] \\ &= -\sum_{i=1}^{n}\left(X_{ia}X_{ib}\pi_i\left(1-\pi_i\right)\right).\end{aligned} \tag{2.37b}$$

We may write the derivative equations in matrix notations to simplify the calculations. Refer to matrix approach to multiple logistic regression model and let **u** be the column vector of ones, $\mathbf{u}'=[1,1,...,1]$, the first derivative of log-likelihood function can be written in matrix notation as follow.

$$\frac{\partial \ln L(\beta)}{\partial \beta_0} = \sum_{i=1}^{n} (Y_i - \pi_i) = \mathbf{u}'(\mathbf{Y} - \boldsymbol{\pi}) \tag{2.38a}$$

$$\frac{\partial \ln L(\beta)}{\partial \beta_a} = \sum_{i=1}^{n} X_{ia}(Y_i - \pi_i) = \mathbf{X}'(\mathbf{Y} - \boldsymbol{\pi}) \tag{2.38b}$$

Notice that the first column of data matrix **X** is consist of ones which denotes the regression constant β_0. This notation provides us to combine the equations (2.38a) and (2.38b). Thus (2.38a) and (2.38b) have form

$$\frac{\partial \ln L(\beta)}{\partial \beta} = \sum_{i=1}^{n} X_i(Y_i - \pi_i) = \mathbf{X}'(\mathbf{Y} - \boldsymbol{\pi}). \tag{2.39}$$

From (2.37a) and (2.37b), we define **W** be the diagonal matrix with elements $\pi_i(1-\pi_i)$. Similar to equation (2.39), the second derivatives are rewritten as follow.

$$\frac{\partial^2 \ln L(\beta)}{\partial \beta_a \beta_b} = -\sum_{i=1}^{n} \left(X_{ia} X_{ib} \pi_i (1-\pi_i)\right) = -\mathbf{X}'\mathbf{W}\mathbf{X} \tag{2.40}$$

The matrix of first derivates, $\mathbf{X}'(\mathbf{Y} - \boldsymbol{\pi})$, is called *Gradient matrix*. Similarly, the matrix of second derivates is called *Hessian matrix*. The ML estimates of unknown parameters have a large-sample normal distribution. The estimated variances of regression parameters are equal to the inverse of the information matrix. The observed information matrix has elements;

$$\mathrm{E}\left[\left(\frac{\partial l(\boldsymbol{\beta})}{\beta_j}\right)^2\right] = -\mathrm{E}\left[\left(\frac{\partial^2 l(\boldsymbol{\beta})}{\beta_a \beta_b}\right)\right] = -\left(-\sum_{i=1}^{n}\left(X_{ia}X_{ib}\pi_i(1-\pi_i)\right)\right) = \mathbf{X'WX}. \tag{2.41}$$

The estimated covariance matrix is equal to

$$\mathrm{cov}\left(\hat{\boldsymbol{\beta}}\right) = \{\mathbf{X'WX}\}^{-1} \tag{2.42}$$

where the square roots of the main diagonal elements of (2.42) are equal to the standard error of $\hat{\boldsymbol{\beta}}$.

Substituting (2.39) and (2.40) into NR equation (2.36), the iterative equation has the form

$$\hat{\beta}^{t+1} = \hat{\beta}^t - \frac{\partial l(\beta^t)/\partial\beta}{\partial^2 l(\beta^t)/\partial\beta^2} = \hat{\beta}^t - \frac{\mathbf{X'(Y-\hat{\pi})}}{-\mathbf{X'WX}}$$

$$\hat{\boldsymbol{\beta}}^{t+1} = \hat{\boldsymbol{\beta}}^t + (\mathbf{X'WX})^{-1}\mathbf{X'(Y-\hat{\pi})}. \tag{2.43}$$

With an successful initial guess $\hat{\boldsymbol{\beta}}^{(0)}$, the iterations continue until the function converges. At final iteration, the parameter estimations converge to the ML estimates. The selection of initial guess has an influence on the number of iterations. In literature there are several ways suggested how to select the initial guess. We will focus on this issue in the following chapters.

CHAPTER THREE

PENALIZED LOGISTIC REGRESSION

In chapter two, we focused on modeling the logistic regression and estimating the unknown parameters. Logistic regression is widely used statistical tool for modeling the main effects and interactions for a binary response variable. However, for some data sets, logistic regression models might have considerable drawbacks that reduce the fitting performance of the models. We will mention these drawbacks in the following sections. Since model fitting is an important part of all statistical applications, we should overcome these drawbacks. We can modify the logistic regression model to eliminate these problems.

In this chapter, we will study the Penalized Logistic Regression (PLR) as a modification of logistic regression model to overcome the unexpected problems. Penalized logistic regression has been proposed especially for cancer classifications. The rising of microarray applications has enabled the researchers to measure the thousands of interactions between DNA sequences simultaneously. Recent works in microarray applications have shown that most of the common diseases are highly correlated with gene interactions. These gene interactions can be used to classify the different types of tumors and other genetic diseases. The size of DNA sequences, say billions of gene sequences and interactions, makes the PLR method one of the most important statistical tools for modeling microarray data sets because modeling the large samples is harder.

Many recent articles that concern with genetic variations have been published for classifying cancer using the DNA information. Such as the articles Golub *et al.* (1999), Zhu and Hastie (2004). Although PLR is mostly called with medical applications, it can be used for different types of data sets.

3.1 Penalizing the Model

The size of the data set may vary with the scope of application. Microarray applications have large data sets because DNA sequences have billions of gene interactions. While working with DNA applications, the gene and its interactions are defined as explanatory variables. We can see that there is huge number of explanatory variables *k*. However, the number of observations *n* is small comparing to the number of predictors. There are several drawbacks in this situation:

- Because $k >> n$, there will be more unknown than equations. The feasible solutions will be infinite.
- The regression model may have overfitting problem. For a training set we may have zero prediction error, i.e. the model fits perfectly for training set. However, for a new sample, the prediction might be very poor.
- The multicollinearity may exist among the predictors. Hence, the estimated parameters may not be reliable.

Because of these problems, we should modify the corresponding regression model. These problems can be solved by penalizing the logistic regression formulation. The penalization term is added into likelihood equation and the parameter estimations are made by using the penalized likelihood equation. Let $J(\beta)$ be the penalty term and L^* be the penalized log-likelihood equation, we may define L^* as follow:

$$L^* = l(\beta) - \frac{\lambda}{2} J(\beta) \tag{3.1}$$

where $l(\beta)$ denotes the Bernoulli log-likelihood and λ denotes the penalty parameter. The parameter estimate not only depends on penalty parameter λ but also penalization term $J(\beta)$. The penalty term $J(\beta)$ has the following form:

$$J(\beta) = \sum_k \gamma_k \psi(\beta_k) \quad , \quad \gamma_k > 0 \,. \tag{3.2}$$

In literature there are several penalization criteria. The selection of inner function $\psi(\beta_k)$ specifies the penalization method. Each penalization method might slightly change the parameter estimations. Thus, selecting the penalization method is significantly affects the parameter estimates.

Quadratic penalization or ridge regularization (L_2) and LASSO penalization (L_1) are commonly used penalization methods for logistic regression. In this paper, we use quadratic penalization which is frequently preferred for microarray and medical applications. This penalization technique often works pretty well. In literature, several penalty functions have been proposed. Table 3.1 gives a small list of penalty functions.

Tablo 3.1 Penalization Functions (Antoniadis, A., 2003)

Penalty Function	Author
$\psi(\beta) = \lvert\beta\rvert^{\gamma}$	Saquib(1998)
$\psi(\beta) = \gamma\lvert\beta\rvert / (1 + \gamma\lvert\beta\rvert)$	Geman(1992,1995)
$\psi(\beta) = \gamma\beta^2 / (1 + \gamma\beta^2)$	McClure(1987)

3.1.1 Quadratic (L_2) Penalization

Assume that we have a sample with *n* observations and *k* predictors. The L_2 penalization is based on the norm of the predictor vector, that is:

$$J(\beta) = \sum_k \gamma_k \psi(\beta_k) \\ = \|\boldsymbol{\beta}\|^2 = \sum_k \beta_k^2 . \tag{3.3}$$

Substituting (3.3) into likelihood function (3.1), the penalized log-likelihood function equals to:

$$L^* = l(\beta) - \frac{\lambda}{2} \sum_k \beta_k^2 . \tag{3.4}$$

The ML method estimates the penalized logistic regression parameters in the same manner as logistic regression parameters by maximizing (3.4). The regression parameters are estimated with Newton-Raphson (NR) algorithm. We have studied the NR algorithm and the mathematical calculations in chapter 2. To apply NR algorithm (2.43), we should calculate the first and second derivatives of penalized log-likelihood function. The first and second derivatives of (3.4) are:

$$\frac{\partial L^*}{\partial \beta_k} = \sum_{i=1}^{n} Y_i X_{ik} - \sum_{i=1}^{n} \left(\frac{X_{ik} \exp\left(\beta_0 + \beta_1 X_{i1} + \cdots + \beta_{p-1} X_{i(p-1)}\right)}{1 + \exp\left(\beta_0 + \beta_1 X_{i1} + \cdots + \beta_{p-1} X_{i(p-1)}\right)} \right) - \lambda \beta_k = \mathbf{X'}(\mathbf{Y} - \hat{\boldsymbol{\pi}}) - \lambda \boldsymbol{\beta} \tag{3.5}$$

$$\frac{\partial L^*}{\partial \beta_a \partial \beta_b} = -\sum_{i=1}^{n} \left(X_{ia} X_{ib} \pi_i (1 - \pi_i) \right) - \lambda = -\mathbf{X'WX} - \lambda \mathbf{I} \tag{3.6}$$

Substituting (3.5) and (3.6) into NR formulation, we repeat NR steps to obtain the following parameter estimates:

$$\hat{\boldsymbol{\beta}}^{t+1} = \hat{\boldsymbol{\beta}}^{t} - \frac{\partial L^{*} / \partial \boldsymbol{\beta}}{\partial^{2} L^{*} / \partial \boldsymbol{\beta}' \partial \boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}^{t} - \frac{\mathbf{X}'(\mathbf{Y} - \hat{\boldsymbol{\pi}}) - \lambda \boldsymbol{\beta}}{-\mathbf{X}'\mathbf{W}\mathbf{X} - \lambda \mathbf{I}}$$

$$\hat{\boldsymbol{\beta}}^{t+1} = (\mathbf{X}'\mathbf{W}\mathbf{X} + \boldsymbol{\Lambda})^{-1}\mathbf{X}'\mathbf{W}\left\{\mathbf{X}\hat{\boldsymbol{\beta}}^{t} + \mathbf{W}^{-1}(\mathbf{Y} - \hat{\boldsymbol{\pi}})\right\}. \tag{3.7}$$

where $\boldsymbol{\Lambda}$ is $(p \times p)$ diagonal matrix with elements $[0, \lambda, \ldots, \lambda]$. (3.7) is equivalent to *iteratively reweighted algorithm* form. Notice that the first element of $\boldsymbol{\Lambda}$ is zero that means no penalization on regression constant. Equation (3.7) can be solved iteratively. The iteration starts with an initial guess. In most applications the function converges within ten iterations. Possible starting values for parameters are $\hat{\beta}_0 = \log[\bar{y}/(1-\bar{y})]$ with $\bar{y} = \sum_{i=1}^{n} y_i / n$ and $\hat{\boldsymbol{\beta}} = \mathbf{0}$.

3.1.2 Advantages of Quadratic Regularization

Using quadratic penalization improves the performance of the logistic regression model. When the number of explanatory variables is large, the prediction error of the observations becomes smaller than it should be. The prediction error of the *i*th observation is:

$$\begin{aligned} e_i &= Y_i - \hat{\pi}_i \\ &= Y_i - \frac{\exp\left(\beta_0 + \sum_{j=1}^{k} x_{ij}\beta_j\right)}{1 + \exp\left(\beta_0 + \sum_{j=1}^{k} x_{ij}\beta_j\right)}. \end{aligned} \tag{3.8}$$

Notice that large number of explanatory variables for (3.8) leads to serious overfitting problem. The SSE of the model becomes small for training set but large for new samples. However, the L_2 penalization controls the magnitude of regression parameters. We can see from (3.4) that the penalization parameter shrinks the

regression parameters. The larger the λ, the smaller the β_j's are forced to be. Thus, the penalization provides us to fit the regression model in a stable fashion since the regularization parameter controls the regression parameters.

When we include high-order interaction terms in the model the number of predictors grows rapidly with the possible result of multicollinearity amongst the predictors. The multicollinearity induces infinite solution for unknown parameters. The quadratic penalization easily overcomes the multicollinearity problem by penalizing the regression parameters. Thus, unique solution is obtained for regression parameters. Overfitting and multicollinearity problem yields very unstable estimates for the regression model. One should handle this problem carefully before using regression model on the new samples.

Logistic regression models may be applied into grouped or ungrouped data sets. When the observations are grouped, some of the cells in the contingency table might be very small, say smaller than 5, or zero. Zero cells in contingency tables lead to poor estimates. We cannot fit a logistic regression model with zero cells in contingency tables. However, when the model is modified with quadratic penalization, the PLR model fits the observations with zero cells. The corresponding regression coefficients of empty cells will be automatically set to zero. When the observations are grouped, the binomial distribution is appropriate to fit the logistic regression model. Using the binomial log-likelihood the ML estimates are obtained. However, we should consider the symmetric constraint on the regression coefficients, that is, $\sum_k \beta_k = 0$. This constraint is automatically satisfied when the quadratic penalization is applied to the multinomial likelihood function.

As we already discussed in the earlier sections, the PLR method fits the regression model by eliminating the multicollinearity amongst the predictors. Notice that PLR works with the original data and variables to estimate the regression parameters

without multicollinearity by maximizing the penalized log-likelihood or minimizing the residual sum of squares, respectively. Thus, we obtain parameter estimate for each explanatory variables which may provide us to measure the effect of each predictors on the success probability.

3.2 Choosing the Regularization Parameter

The choice of regularization parameter is crucial since the estimates of the regression coefficients are affected by λ. There are several measures to determine the best choice of regularization parameter. These measures mostly based on two popular data-driven techniques. The first one is based on Cross-Validation (CV) and the other on the Akaike Information Criterion (AIC). We use CV based measures to define the best value of regularization parameter. If the regularization parameter is selected correctly, the prediction error of the PLR model becomes smaller. Note that determining the regularization parameter by considering only one measure may not be reasonable. We should check several measures simultaneously to obtain the best value of λ.

Suppose we have sample with *n* paired sets (y_i, x_i) and *k* predictors. Cessie and Houwelingen (1992) concentrate on three different measures to quantify the prediction error of the model. These measures are:

(1) Classification or Counting Error (CE)

$$CE = \begin{cases} 1 & \text{if} \quad Y_i = 1 \quad \text{and} \quad \pi_i^{\lambda} < \dfrac{1}{2} \\ & \text{or} \quad Y_i = 0 \quad \text{and} \quad \pi_i^{\lambda} > \dfrac{1}{2} \\ \dfrac{1}{2} & \text{if} \quad \pi_i^{\lambda} = \dfrac{1}{2} \\ 0 & \text{otherwise} \end{cases} \tag{3.9}$$

(2) Squared Error (SE)

$$SE = (Y_i - \hat{\pi}_i^{\lambda})^2 \tag{3.10}$$

(3) Minus log-likelihood Error (ML)

$$ML = -\left\{Y_i \log \hat{\pi}_i^{\lambda} + (1 - Y_i) \log(1 - \hat{\pi}_i^{\lambda})\right\} \tag{3.11}$$

where $\hat{\pi}_i^{\lambda}$ denotes the penalized probability that $Y_i = 1$, i is for i-th case in the new sample. The magnitude of three measures (1-3) becomes maximum while the real success probability is around 0.5, that is $\pi = 0.5$. Note that the measures (1-3) are calculated from new sample by using the penalized parameter estimates obtained from training set. Thus, small values for 1,2 and 3 indicate the successful classification on new samples. When a validation set (new sample) is available, the predicted values of new sample can be compared for various values of λ and the optimal solution for regularization parameter is achieved. However, if a validation set is not available, the cross-validated measures can be used. Cross-validation predicts each observation based on other observations. Let $\beta_{(-i)}^{\lambda}$ be the parameter estimate based on all observations except the i-th observation and $\hat{\pi}_{(-i)}^{\lambda}$ be the predicted probability of removed observation obtained from the equation based on $\beta_{(-i)}^{\lambda}$'s. Then, the cross-validated prediction errors of three measures defined above are (Cessie and Houwelingen (1992)):

(1) Mean Classification or Counting Error (MCE)

$$MCE_{CV} = n^{-1} \sum_{i=1}^{n} \left[Y_i [\hat{\pi}_{(-i)}^{\lambda} < \frac{1}{2}] + (1 - Y_i)[\hat{\pi}_{(-i)}^{\lambda} > \frac{1}{2}] + \frac{1}{2}[\hat{\pi}_{(-i)}^{\lambda} = \frac{1}{2}] \right]. \tag{3.12}$$

The expressions inside the brackets are equal to $[.]=1$, if it is true and $[.]=0$ if it is false.

(2) Mean Squared Error (MSE)

$$MSE_{CV} = n^{-1}\sum_{i=1}^{n}\left\{(Y_i - \hat{\pi}_{(-i)}^{\lambda})^2\right\}. \tag{3.13}$$

MSE_{CV} is also known as PRESS (Prediction Residual Sum of Squares) statistics.

(3) Mean Minus log-likelihood Error (MLL)

$$MML_{CV} = -n^{-1}\sum_{i=1}^{n}\left\{Y_i \log \hat{\pi}_{(-i)}^{\lambda} + (1-Y_i)\log(1-\hat{\pi}_{(-i)}^{\lambda})\right\}. \tag{3.14}$$

Two more extensions of (3.13) are available in Cessie and Houwelingen (1992). The effect of influential and outlier observations is introduced to the cross-validated MSE measure. In addition to (1,2, and 3), we may also use alternative measures. Let L_i^{λ} be the penalized log-likelihood for *i*th observation based on parameter estimates β_i^{λ} which is estimated from entire data and $L_{(-i)}^{\lambda}$ be the penalized log likelihood when *i*th observation is removed which is based on $\beta_{(-i)}^{\lambda}$. Using the same analogy from (3.13) and substituting into log-likelihood function, we get SSLE (Sum of Squared log-likelihood Error):

$$SSLE = \sum_{i=1}^{n}(L_i^{\lambda} - L_{(-i)}^{\lambda})^2 \tag{3.15}$$

where

$$L_i^{\lambda} = Y_i \log \hat{\pi}_i^{\lambda} + (1 - Y_i) \log(1 - \hat{\pi}_i^{\lambda}) \tag{3.16}$$

$$L_{(-i)}^{\lambda} = Y_i \log \hat{\pi}_{(-i)}^{\lambda} + (1 - Y_i) \log(1 - \hat{\pi}_{(-i)}^{\lambda}) . \tag{3.17}$$

The best choice of regularization parameter according to (3.15) is the point that minimizes the corresponding function.

CHAPTER FOUR

NUMERICAL STUDY

In statistical researches, the scope of the study and the properties of the data set significantly affect the selection of analyzing methods. We can analyze data by using several statistical techniques. Equivalently, several statistical models can be conducted on the same data set. However, only very few of these models provide us statistically significant results. Thus, the determination of best model is crucial. We should carefully understand the data and the model assumptions to ensure the best selection of statistical modeling techniques.

Logistic regression is a special form of General Linear Models (GLMs) which is commonly used for binary response variables. Although it's mostly used with binary responses, we can use LR models with multi-category responses (more than two categories). The categories might be either in ordinal scale or not. We can select the best criteria with respect to properties of response variable. For all cases, the LR method models the success probabilities of dependent variable.

In practice, the data might be stored in different forms. The sample observations might be grouped or ungrouped. When the data is grouped, we use contingency tables to predict the cell probabilities. Thus, the response variable is distributed as Binomial. The regression parameters are estimated by using the Binomial likelihood equations.

In this study, we use coronary heart disease data to compare the LR and PLR model. The data is stored as ungrouped.

4.1 Coronary Heart Disease Data

Let Y be a binary response variable with categories 0 and 1 for coronary heart attack condition. The patient is assigned to class 1 if have heart attack. The sample has 16 patients where 9 have had heart attack. There are 22 explanatory variables including the physical measures (weight, obesity, age, gender etc.) and blood test results (HDL, LDL, Blood Pressure, Platelet, Cholesterol etc.). The data has multicollinearity and high dimension problem. The data is given in Appendix.

We fit LR and PLR models to data. Table 4.1 gives the basic results of LR analysis.

Table 4.1 Classification rate of LR model

Observed			Predicted		
			Heart Attack		Percentage Correct
			0	1	%
Step 0	Heart Attack	0	0	7	.0
		1	0	9	100.0
	Overall Percentage				56.3

a Constant is included in the model.
b The cut value is 0.50

As it's seen from Table 4.1, all patients are assigned to group 1. All observations from class 0 are misclassified. The LR model has very low classification rate with 56,3%. Remember that the data has multicollinearity and high dimension problems, the LR model cannot estimate the regression coefficients or the estimated coefficients will be insignificant. The regression function includes only constant term which leads the estimated probabilities to be larger than 0.5. Table 4.2 gives the significant test results for predictors which are not included in the regression function.

Table 4.2 Significance test of variables not in the model

	VAR002	**VAR003**	**VAR004**	**VAR005**	**VAR006**	**VAR007**	**VAR008**	**VAR009**
Coeff.	0.00403	2.10409	0.02657	0.4232	0.00403	2.44643	0.05162	2.28571
Significance	0.94937	0.14691	0.87052	0.5153	0.94937	0.11779	0.82026	0.13057

	VAR010	**VAR011**	**VAR012**	**VAR013**	**VAR014**	**VAR015**	**VAR016**	**VAR017**
Coeff.	0.26907	0.00090	0.37299	1.84780	6.16683	1.26101	0.03872	1.24336
Significance	0.60396	0.97610	0.54138	0.17404	0.01302	0.26146	0.84400	0.26483

	VAR018	**VAR019**	**VAR020**	**VAR021**	**VAR022**	**VAR023**
Coeff.	0.57838	3.53556	3.26361	2.13741	0.56341	0.00150
Significance	0.44695	0.06007	0.07083	0.14374	0.45289	0.96912

Because of the high dependency among the predictors, the standard errors for regression coefficients are to be overestimated. This problem yields the predictors become insignificant. Thus, the LR model is not a suitable method for coronary heart attack data.

We fit PLR model based on quadratic penalization to the same data set. The variable selection and significance test of coefficients are excluded for PLR model. Our primary concern is to focus on the parameter estimation steps of PLR method. We developed the PLR codes with MATLAB (ver. 7.7.0.471), and the parameter estimates are obtained with Newton-Raphson algorithm by using these codes. These codes are given in Appendix. Before we estimate the regression coefficients, the optimum value of penalty parameter should be determined. Figure 4.1 gives the cross validated errors to determine the optimum regularization parameter. The predicted errors are rescaled to combine three measures in the same graph.

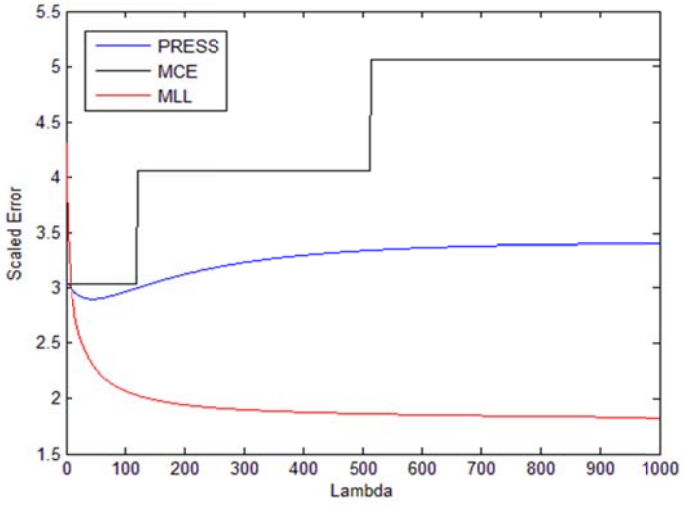
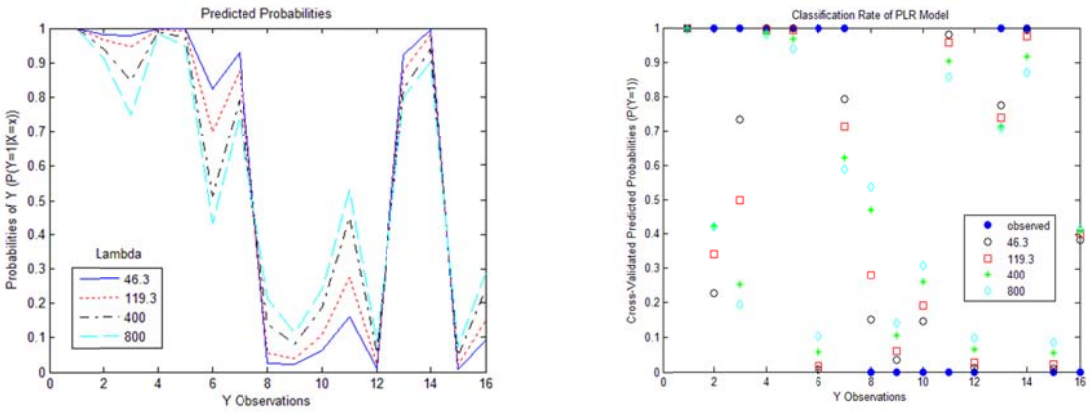


Figure 4.1 Cross-validated errors of PLR for various penalty parameters.

From Figure 4.1, minimum PRESS is obtained when penalty parameter is selected 46.3. Similarly, the minimum MCE is obtained while penalty parameter is selected in the interval [0, 119.3]. Unlikely MCE and PRESS, the MLL measure is decreasing function. The error decreases very slowly when penalty parameter exceeds 150.



(a) (b)

Figure 4.2 Predicted probabilities: (a) including all observations, (b) LOO cross-validated

Figure 4.2 shows the predicted probabilities of PLR model when all observations are in the model, (a), and when leave-one-out cross-validation is performed, (b). While the penalty parameter increases, the success probabilities move towards the threshold point 0.5. Thus the prediction error increases for increasing values of regularization parameter. The blue filled circles indicate the sample observations. Observation 6 and 11 are misclassified at all levels of penalty parameter.

The number of iterations for NR algorithm can be controlled with tolerance level $\left\|\boldsymbol{\beta}^{t+1}-\boldsymbol{\beta}^{t}\right\|^{2}$. When tolerance level is selected very small, the number of iterations and the elapsed time for convergence grow rapidly. In addition, the tolerance level may significantly affect the amount of cross-validated prediction errors. We employ NR codes with different tolerance levels. Figure 4.3 shows the effect of tolerance level on the predicted errors and determination of penalty parameter.

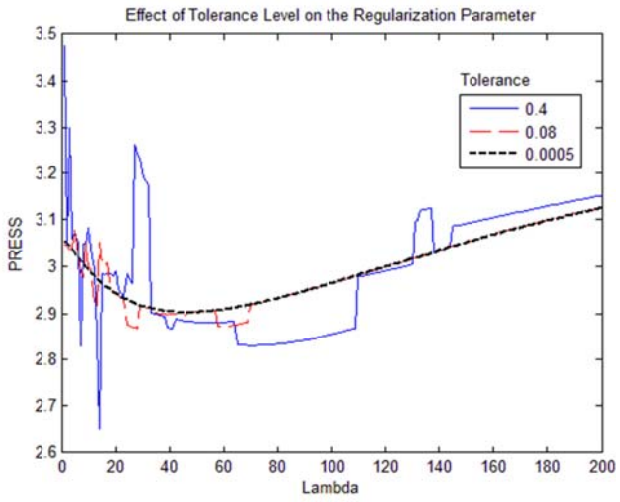


Figure 4.3 Effect of tolerance level on PRESS

The predicted errors instantly increases or decreases if the tolerance level is selected large. The larger the tolerance level, the smoother the error lines to be. In this study, we selected tolerance level as 0.0005.

4.2 Comparison of LR and PLR models

According to three cross-validated measures and different penalty parameters, we calculated the classification rate of PLR and LR models. Table 4.3 gives the classification rate of PLR for training set and LOO cross-validated samples.

Table 4.3 Misclassification rate of PLR model

	$\lambda = 46.3$	$\lambda = 119.3$	$\lambda = 400$	$\lambda = 800$
Cross-Validated (PLR)	3/16	4/16	4/16	5/16
No Validation (PLR)	0/16	0/16	0/16	2/16

Finally, we compare classification rate of LR and PLR model with Table 4.4.

Table 4.4 Misclassification rate of LR and PLR model

	PLR ($\lambda = 46.3$)	LR
Misclassification Rate	3/16	7/16

From tables 4.3 and 4.4, the minimum misclassification rate is obtained when the penalty parameter is selected as 46.3. The overall classification rate for PLR is 13/16 (81.25%). Thus, we select the penalty parameter as 46.3. At this point, we can estimate the penalized regression coefficients.

Table 4.5 Penalized regression coefficients.

	Constant	**Gender**	**Age**	**Weight**	**Obesity**	**Smoking**	**Systolic Blood P.**	**…**
λ = 0.005	-29.5735	0.006737	0.289104	-0.03655	-0.00374	0.005113	0.137307	…
λ = 46.3	-10.6119	0.001018	0.061133	-0.00172	-7.0E-05	0.000461	0.026749	…

From table 4.5, the regression equation for binary response might be written as it's given with equation (4.1).

$$\hat{Y} = -10.6119 + 0.001018 * Gender + 0.061133 * Age - 0.00172 * Wheight + \ldots \quad (4.1)$$

As it's clearly seen from table 4.5, the regularization parameter shrinks the regression coefficients towards zero. However, none of the coefficients becomes exactly zero. The corresponding coefficient for obesity is estimated very small near zero. This property might be noted as a disadvantage of quadratic penalization since all coefficients are included in the model even if the estimated coefficient is very small. Thus, any predictor will be included in the model even if it's statistically insignificant.

CHAPTER FIVE

DISCUSSION AND FUTURE WORK

In statistical researches, the scope of the study and the properties of the data set significantly affect the selection of analyzing methods. We can analyze data by using several statistical techniques. However, only very few of these models provide us statistically significant results. Thus, the determination of the best model is crucial. We should carefully understand the data and the model assumptions to ensure the best selection of statistical modeling techniques.

Logistic regression is a powerful modeling tool when the response variable is categorical. This method is highly affected by multicollinearity and the size of samples. Thus, we should be cautious against these problems during the analysis. We can eliminate these problems by employing alternative methods or introducing some useful modifications into logistic regression analysis. In this thesis, we prefer to modify logistic regression for better results rather than changing the modeling technique.

The data set which was used in numerical study had multicollinearity and high dimension problem which led the method to be inaccurate for fitting the regression line to the data. The response variable had only two categories. We proposed PLR model to eliminate these problems. This study focused on determination of penalization criteria and the parameter estimation of PLR method. Since PLR method does not remove the multicollinearity problem from data set, the high dependency still exists amongst the predictors. From earlier chapters, it can be concluded that PLR method easily overcome multicollinearity problem without removing it and estimates robust regression coefficients against the multicollinearity problem.

In chapter three, we have focused on quadratic penalization for PLR method. We have addressed some important implications about quadratic penalization. The regression coefficients are estimated at the optimum level of penalty parameter. In order to define the optimum penalty parameter, we examined several measures simultaneously to minimize the regression error. Each measure gives its minimum at different levels of lambda. The overall classification rate is calculated for different levels of penalty parameter. Finally, we define the best choice of penalty parameter with respect to classification rate and prediction error. Moreover, advantages and disadvantages of quadratic penalization were also provided. One important point of quadratic penalization is that the regression coefficients never estimated as zero. Although the penalty parameter shrinks regression coefficients towards zero, they never become exactly zero.

According to the results of the numerical study, the overall classification rate for LR model increased significantly when the PLR model was performed. While some coefficients were estimated very small, none of the regression coefficients were estimated exactly zero. Although the overall classification rate was quite high, the regression model was not defined as the best model. Because there were many predictors, some of them might be redundant. We might remove redundant predictors by performing variable selection methods. However, the variable selection criterion is not the scope of this study. In addition, an alternative penalization method might be performed on the same data set to compare the performance of quadratic penalization. For example, LASSO penalization is also popular method which is used with logistic regression. Unlike quadratic penalization, LASSO is able to shrink regression parameters exactly to zero. This property provides us an important advantage. LASSO penalization internally performs variable selection process. We may select only required predictors by removing the regression predictors where the estimated coefficients equal zero.

In recent researches, LR method is frequently used with microarray data analysis. Since there is high dependency among the DNA sequences, the penalization techniques are considered with LR model for better classification. In this thesis, we performed quadratic penalization on the coronary heart attack data set. For further research, we will consider quadratic penalization with microarray data for Alzheimer disease. The data have gene sequences for 1305 patients. For each patient, more than 620000 genes are observed. As a comparison of penalization methods, we will perform quadratic and LASSO penalization on the same data set.

REFERENCES

Agresti, A. (2002). *Categorical data analysis* (Second Ed). NY: Wiley.

Agresti, A. (2007). *An introduction to categorical data analysis* (2nd Edition). NY: Wiley.

Aguilera, Ana M., Escabias, M. & Valderrama, Mariano J. (2006). Using principal components for estimating logistic regression with high-dimensional multicollinear data. *Computational Statistics & Data Analysis*, 50, 1905-1924.

Antoniadis, A. (2003). Penalized logistic regression and classification of microarray data. *Laboratoire IMAG-LMC*, University Joseph Fourier, France, Powerpoint Presentation.

Cessie, S. Le & Van Houwelingen, J. C. (1992). Ridge estimator in logistic regression. *Applied Statistics*, 41, No.1, 191-201.

Fleming, W. (1977). *Functions of several variables* (2nd Edition). NY: Springer-Verlag

Golub, T., Slonim, D., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J., et al. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286, 531-536.

Hawkins, Douglas M. (2004). The problem of overfitting, *J. Chem. Inf. Computational Sci.,* 44, 1-12.

Hoerl, Arthur E. & Kennard, Robert W. (2000). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics,* 42, No.1.

Neter, J., Kutner, M. H., Nachtsheim, Christopher J., Wasserman, W. (1996). *Applied linear statistical models* (4th Edition). USA: Irwin.

Park, M. Y., Hastie, T. (2007). Penalized logistic regression for detecting gene interactions. *Oxford Journals*, 9, 30-50.

Shen, L., Tan, Eng C. (2005). PLS and SVD based penalized logistic regression for cancer classification using microarray data, in *Third Asia-Pacific Bioinformatics Conference* (APBC2005) (Chen, P. and Wong, L., Eds.), 219-228, Imperial College Press, Singapore.

Zhu J., Hastie T. (2004). Classification of gene microarrays by penalized logistic regression. *Biostatistics*, 5, No.3, 427-443.

APPENDIX

A. NR codes for MATLAB

```
%
clc
display('Before running these codes, the data matrix should be denoted as X and reponse vector as Y (with CAPITAL letters)')
inc = input('Define the increasements for penalty parameter  :  ');  % the loop repeats with the steps of inc for penalization parameter.
lambda_lower = input('Lower limit of lambda (starting value for the loop of penalization)  :  ');
lambda_upper = input('Upper limit of lambda :  ');
tolerance = input('Define the tolerance level for convergence :  '); % The tolerance level affects the number of iterations for NR algorithm.
%%%
[n,p] = size(X);  % The data matrix X should be consisted of column of ones in the first column to estimate the regression constant b0.
Y_out = [];
X_out = [];
logit_out = [];
resultd_i = [];
RESULTD = [];
pi_hatd = [];
k = 100;
mlambda = [];
RESD_M = [];
PRESSD_M = [];
LOGL = [];
CEcond = [];
Y_CE = [];
CE_out = [];
CE = [];
lambda_RESULT = [];
%
%
for lambda=lambda_lower:inc:lambda_upper % the loop for penalty parameter, lambda.
%
 mlambda = lambda*eye(p);
 mlambda(1,1) = 0;
%
 LOGL = [];
 Xd = [];
 Yd = [];
```

```
logitd = [];
%
for d=1:n   % the loop for Cross-Validation
    Xd = [X(1:d-1,:);X(d+1:n,:)];
    Yd = [Y(1:d-1,:);Y(d+1:n,:)];
    Y_out = Y(d,:);
    X_out = X(d,:);
    b0d_initial = log(mean(Yd)/(1-mean(Yd)));
    betad_initial = [b0d_initial;zeros(p-1,1)];  % Cross-Validated initial guesses for
coefficients which is used for the starting value of NR algorithm.
    betad = betad_initial;
    kd = 100;

while kd > tolerance  % Loop for NR algorithm (Cross Validated.)

   pi_hatd = [];
   logitd = Xd*betad;  % column vector of logits.

  for i=1:n-1  % loop for obtaining the W matrix.
     prob = exp(logitd(i,1))/(1+exp(logitd(i,1)));
     pi_hatd = [pi_hatd;prob];
     Wd(i,i) = prob*(1-prob);
  end % end of loop for W matrix

   betad_new = (inv(Xd'*Wd*Xd + mlambda))*Xd'*Wd*(Xd*betad +
inv(Wd)*(Yd-pi_hatd));
   kd = norm(betad_new-betad);
   betad = betad_new;

end % end of loop for NR algorithm

   logit_out = X_out*betad;  % logit of removed observation.
   Y_hat_out = exp(logit_out)/(1+exp(logit_out)); % estimated probability of
removed observation.
   logL_out = Y_out*log(Y_hat_out)+(1-Y_out)*log(1-Y_hat_out);%log-likelihood
value of removed observation
   %
   %%%% MCE (determination of the conditions in the brackets for MCE measure.)
      if Y_hat_out < 0.5
            CEcond(d,:) = [1,0,0];
         elseif Y_hat_out > 0.5
            CEcond(d,:) = [0,1,0];
         else
            CEcond(d,:) = [0,0,1];
      end
   %
   Y_CE = [Y_out;1-Y_out;0.5];
   CE_out = CEcond(d,:)*Y_CE;
```

```
    CE = [CE;CE_out];
    %
    %%%% end for MCE calculation.
    %
    LOGL = [LOGL;logL_out];
    RESD = Y_out-Y_hat_out; % residul for ith observation.
    RESD_M = [RESD_M;RESD];   % residuals vector.
    PRESSD = (norm(RESD_M))^2;

 end  % end for loop Cross-Validation

    MLL = (-1/n)*sum(LOGL);
    MCE =(1/n)*sum(CE);
    RESULTD_i = [lambda;PRESSD;MCE;MLL];
    RESULTD = [RESULTD RESULTD_i];
    RESD = [];
    RESULTD_i = [];
    RESD_M = [];
    PRESSD = [];
    CE = [];

end % end of loop for lambda (penalization)
%
[MinPRESS,index1] = min(RESULTD(2,:));
MinPRESS_Lambda = (RESULTD(1,index1));
[MinMCE,index2] = min(RESULTD(3,:));
MinMCE_Lambda = (RESULTD(1,index2));
[MinMLL,index3] = min(RESULTD(4,:));
MinMLL_Lambda = (RESULTD(1,index3));
%
lambda_RESULT = [MinPRESS,MinPRESS_Lambda
                 MinMCE,MinMCE_Lambda
                 MinMLL,MinMLL_Lambda];
%
%
% when the algorithm converges, the matrix RESULTD gives the selected lambda
values and the calculated error measures for corresponding lambda values. When we
repeat the loop for lambda in a very wide interval, say [1,1000], it might be hard to
find the optimum lambda by skimming the RESULTD matrix. The matrix
lambda_RESULT gives the minimum of these measures from RESULTD and the
corresponding lambda values to achieve these minimum points.
```

B. Coronory Heart Attack Data

Heart Attack	Gender	Age	Weight	Obesity	Smoking	Systolic Blood Pressure	Diastolic Blood Pressure	Hypertension	HDL	LDL	Cholesterol	Triglyceride	Uric Acid	Na	K	CL	Creatine	Platelet	AST (SGOT)	ALT(SGPT)	Alkaline Phosphatase (ALP)	BPM
1	0	56	70	0	1	142	82	1	33	139	257	446	7,1	138	4,1	106	0,8	273	22	25	76	85
1	1	62	75	1	0	126	60	0	40	160	270	460	6,9	136	4,1	100	0,42	111	21	12	109	80
1	0	52	85	1	0	139	96	1	46	154	227	105	5	139	3,7	99	0,94	165	97	30	70	89
1	0	65	90	1	1	149	75	1	23	128	196	157	8,9	133	4,8	101	0,9	415	22	24	78	90
1	1	78	80	0	1	144	65	1	28	130	186	171	9,4	140	3,5	101	1,13	309	30	11	78	81
1	0	78	100	1	0	121	73	0	26	124	207	168	7,2	140	3,6	108	0,89	156	27	19	70	65
1	1	55	66	0	1	129	91	1	38	118	178	74	4,7	143	3,6	103	0,86	281	43	17	86	74
0	0	49	80	0	1	127	73	0	39	186	260	174	3,7	136	4,6	96	0,82	176	19	16	56	79
0	1	61	72	0	0	141	81	1	41	121	191	94	3,1	137	3,9	98	0,79	166	17	21	65	81
0	0	52	84	1	0	119	69	0	28	134	194	141	4,1	136	3,7	99	0,88	192	16	14	81	78
0	0	63	102	1	0	123	77	0	44	124	205	138	3,3	143	3,3	107	1,03	260	15	32	89	77
0	1	57	67	0	1	118	76	0	27	126	190	91	4,4	135	3,7	96	1,09	162	19	34	59	72
1	1	56	71	0	1	135	88	1	36	134	188	104	5,1	141	4,7	101	0,68	314	26	17	71	74
1	0	74	77	0	0	127	63	0	29	145	209	161	3,7	147	3,2	97	1,14	293	32	19	57	69
0	0	63	81	0	1	139	87	1	37	133	195	97	5,7	140	3,9	104	0,81	137	18	22	84	79
0	1	59	63	0	1	123	84	0	29	136	197	168	4,3	137	4,7	98	1,1	171	23	34	72	83

LIST OF TABLES

Page

LIST OF FIGURES

Page