2012

Tuning Parameter Selection in L1 Regularized Logistic Regression

Shujing Shi
Virginia Commonwealth University

Tuning Parameter Selection in L_1 Regularized Logistic Regression

In Partial Fulfillment of the Requirements for the degree

Master of Mathematical Science

Statistics Concentration

By

Shujing Shi

Advisor: Dr. Qin Wang

Assistant Professor

Department of Statistical Sciences and Operations Research

Virginia Commonwealth University

Richmond, VA

December 3, 2012

Table of Contents

List of Tables

Acknowledgment

First and foremost I offer my sincerest gratitude to my supervisor, Dr. Qin Wang, who has supported me though out my thesis with his patience and knowledge. Thank him for introducing this interesting topic to me, providing advice and guidance and patiently correcting my writing time after time. One simply could not wish for a better or friendlier supervisor. I would also like to thank Dr. Edward Boone and Dr. Wen Wan for serving as my committee members and their valuable suggestions.

I would also like to take this opportunity to acknowledge my family and friends. Thank my mother and father, Xiumei Yang and Qingfu Shi, and my brother Shucheng Shi for their endless love and support. Thank my friends Xin Li, Yanhui Han, Tian Wu, Jia Zhang and Dengjiang Xing for their encouragement and concerns since the day I arrived this country. They were always there cheering me up and stood by through the good times and bad.

Finally I would like to thank my dearest uncle and aunt, Sixin Yang and Yu Huang, and my cousins, Charles and Emilie Yang for giving a lovely home and family here. They are always so kind to me and I cannot imagine my life in these two years without them.

Abstract

Variable selection is an important topic in regression analysis and is intended to select the best subset of predictors. Least absolute shrinkage and selection operator (Lasso) was introduced by Tibshirani in 1996. This method can serve as a tool for variable selection because it shrinks some coefficients to exact zero by a constraint on the sum of absolute values of regression coefficients.

For logistic regression, Lasso modifies the traditional parameter estimation method, maximum log likelihood, by adding the L_1 norm of the parameters to the negative log likelihood function, so it turns a maximization problem into a minimization one. To solve this problem, we first need to give the value for the parameter of the L_1 norm, called tuning parameter. Since the tuning parameter affects the coefficients estimation and variable selection, we want to find the optimal value for the tuning parameter to get the most accurate coefficient estimation and best subset of predictors in the L_1 regularized regression model.

There are two popular methods to select the optimal value of the tuning parameter that results in a best subset of predictors, Bayesian information criterion (BIC) and cross validation (CV). The objective of this paper is to evaluate and compare these two methods for selecting the optimal value of tuning parameter in terms of coefficients estimation accuracy and variable selection through simulation studies.

CHAPTER 1

Introduction

1.1 Background and Motivation

Variable selection for regression models is a fundamental problem in statistical analysis. We want to select the best subset of predictors that significantly influence the response variables. The popular variable selection methods include but are not limited to the subset selection procedures and information criteria such as Akaike Information Criteria (Akaike, 1973) and Bayesian Information Criteria (Schwarz, 1978). Those methods select predictors in the way that predictors are either retained or eliminated from the model.

In 1996, a shrinkage method called least absolute shrinkage and selection operator (Lasso) was proposed by Tibshirani. The substantial difference between Lasso and the subset selection procedures or the information criteria is that Lasso selects variables and estimates the coefficients simultaneously. Thus, Lasso is a continuous process. Initially, Lasso was proposed for linear regression models and it minimizes the residual sum of squares subject to a L_1 norm constraint, which is the sum of the absolute value of the coefficients being less than a constant (Tibshirani, 1996). The parameter of the L_1 norm constraint is called tuning parameter.

It is important to choose the optimal value of the tuning parameter because it controls the balance of model sparsity and model fitting (Wang et al., 2007). Classical model selection criteria can be used for the selection of the tuning parameter, such as cross validation (CV), AIC and BIC. The performance of BIC and CV for choosing the optimal tuning parameter in terms of

percentage of correctly selected important predictors had been evaluated through simulation studies by Wang et al. (2007). The result shows that overall BIC has higher percentage of correctly selected important predictors than CV. However, the models they setup for the simulation studies are general linear regression models, so their result may not be appropriate to apply to the logistic regression models. Therefore, we are interested that whether BIC would still outperform CV for logistic regression models. The purpose of our study is to examine the performance of the CV and BIC for choosing the optimal value of tuning parameter in terms of variable selection and also estimation accuracy.

CHAPTER 2

Methodology

2.1 Logistic Regression

Logistic regression is widely used to model the outcomes of a binary response. The logistic regression model is a branch of generalized linear models (GLM), a flexible generalization of linear regression that allows for residuals, the difference between the observed response values and the predicted values, not normally distributed. For a binary response, it is not appropriate to use general linear regression because the response values are binomial distributed. The response variable in logistic regression model can also be multi-nominal, taking two or more limited number of possible values. In this work, we focus only on binary outcomes.

All regression models have three components: The response variable $\boldsymbol{Y}$, the linear combination of the predictors, which is the sum of the multiplication of predictors and their coefficients, and the link function, which specifies a function that links the expected value of $\boldsymbol{Y}$ and linear combination of the predictors. The logistic regression model links the linear combination of the predictors with a logit function of the probability of outcome of interest occurring. For a binary response, denote the vector of n response values as $\boldsymbol{Y}$ and its two categories by 1 and 0. Let π be the probability of $y_i = 1$, $(i = 1, 2, \dots, n)$, $\boldsymbol{X}$ be the $n \times (p + 1)$ design matrix, a matrix of predictors with first column being ones, p be the number of predictors and $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \dots, \beta_p)^T$ be the vector of the $p + 1$ parameters corresponding to the design matrix. The logistic regression model has the form

$$logit(\pi) = \ln\left(\frac{\pi}{1-\pi}\right) = \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X} \tag{1.1}$$

, where $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$. The log of the odds, $\ln\left(\frac{\pi}{1-\pi}\right)$, is called the logit transformation of π.

2.1.1 Maximum Likelihood Estimation

Maximum likelihood estimation (MLE) is the standard method of estimating the unknown parameters in a logistic regression model. This method yields values for the unknown parameters which maximize the probability of obtaining the observed response values. The likelihood function expresses the probability of the observed response values as a function of the unknown parameters $\boldsymbol{\beta}$. Let $\pi(x_i; \boldsymbol{\beta})$ be the probability of $y_i = 1$, $(i = 1, 2, \ldots, n)$, since the response variable $\boldsymbol{Y}$ in a two-class logistic regression follows a Bernoulli distribution, the likelihood function is obtained as follows:

$$l(\boldsymbol{\beta}) = \prod_{i=1}^{n} \pi(x_i; \boldsymbol{\beta})^{y_i}[1 - \pi(x_i; \boldsymbol{\beta})]^{1-y_i} \tag{1.2}$$

The log likelihood is defined as the natural log of the equation (1.2):

$$L(\boldsymbol{\beta}) = \ln[l(\boldsymbol{\beta})] = \sum_{i=1}^{n} \{\, y_i \ln \pi(x_i; \boldsymbol{\beta}) + (1 - y_i) \ln[1 - \pi(x_i; \boldsymbol{\beta})]\}$$

$$= \sum_{i=1}^{n} [y_i(\boldsymbol{\beta}^{\mathrm{T}} x_i) - \ln\left(1 + e^{(\boldsymbol{\beta}^{\mathrm{T}} x_i)}\right)] \tag{1.3}$$

To maximize the log likelihood, we set its first derivative to zero. The score equation is

$$\frac{\partial L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{N} x_i\big(y_i - \pi(x_i;\boldsymbol{\beta})\big) = 0 \tag{1.4}$$

The solution to the score equation is the MLE of $\boldsymbol{\beta}$. The Newton-Raphson method can be used to numerically compute the $\boldsymbol{\beta}$, which requires the second-order derivative or Hessian matrix

$$\frac{\partial^2 L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} = -\sum_{i=1}^{n} x_i {x_i}^T \ \pi(x_i;\boldsymbol{\beta})[1 - \pi(x_i;\boldsymbol{\beta})] \tag{1.5}$$

The process of Newton-Raphson method begins with a tentative solution, slightly revises it to check whether it can be improved, and repeats this revision until the improvement is minute, at which point the process is said to have converged (Menard and Scott, 2002).

However, in some instances, the estimation may not reach convergence. When that occurs, the estimated coefficients are meaningless because the iterative process was unable to find the appropriate solution. There might be a number of reasons that cause non-convergence. One reason might be that the ratio of number of predictors and sample size is high. In general, logistic regression models require about 10 observations per predictor (Peduzzi et al., 1996). Another reason might be serious multi-collinearity, which refers to that there are two or more predictors highly correlated. As multi-collinearity increases, standard errors of coefficients estimates increase and the likelihood of model convergence decreases (Menard and Scott, 2002). Separation also might be a reason for non-convergence. Separation occurs when the predictor or a linear combination of several predictors is associated with only one response value when the predictor or the linear combination of several predictors is greater than a constant. Consider the set of data on 10 observations in Table 1. This dataset has a binary response $\boldsymbol{y}$ with value 0 or 1

and one predictor $\boldsymbol{x}$. This is an example of separation problem because when $x < 0$, $y = 0$ and when $x > 0$, $y = 1$.

Table 1 Dataset Exhibiting Separation

$\boldsymbol{x}$	$\boldsymbol{y}$
-5	0
-4	0
-3	0
-2	0
-1	0
1	1
2	1
3	1
4	1
5	1

Figure 1 shows that there are infinite numbers of logistic curves fitted to this dataset, meaning that there are infinite estimates for the coefficients, in which case, the estimation does not reach the convergence.

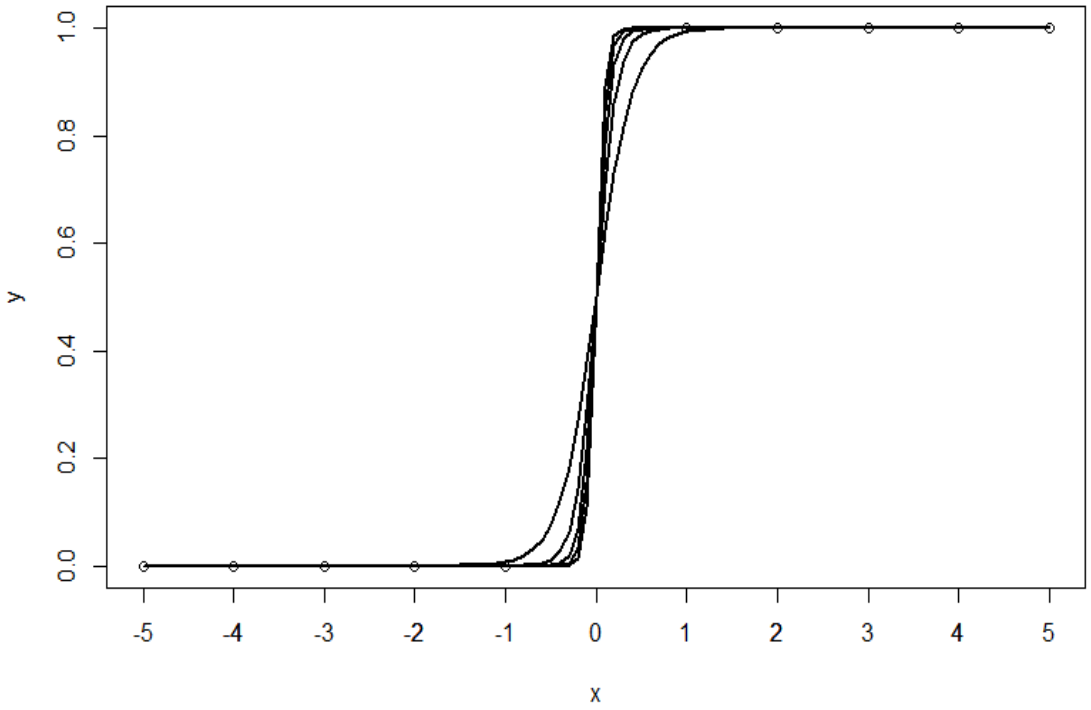

Figure 1 Logistic Curves with Separation Problem

2.1.2 Goodness of Fit

Instead of using R^2 as the statistic for overall fit of a linear regression model, we have deviance for logistic regression. Logistic regression compares the observed values and predicted values based on the log likelihood function defined in equation (1.3). Deviance (D) of a given model is calculated by comparing the given model and the saturated model as expression (1.6). The saturated model contains as many parameters as the sample size so it perfectly fits the data.

$$D = -2 \ln \left(\frac{likelihood\ of\ the\ fitted\ model}{likelihood\ of\ the\ saturated\ model} \right) \quad (1.6)$$

The ratio of likelihood of the fitted model and likelihood of the saturated model in equation (1.6), called likelihood ratio, has a negative value. Model deviance (D) is the multiplication of negative two and the log of the likelihood ratio, which produces an asymptotically chi-squared distributed value. Therefore, upon a chi-square distribution, we can use model deviance as test statistics to assess the model fit. The fit of the model gets poorer as the difference or deviance of the observed values from predicted values gets bigger. The deviance will decrease if we add more variables into the model.

2.1.3 Likelihood Ratio Test

To examine the contribution of individual predictor, we need to test their statistical significance. In linear regression, a coefficient represents the increased value of the predicted response value for each unit increase of that predictor, and the significance of a predictor is assessed by a t test. While in logistic regression, a coefficient represents the increased value of the log odds of probability of occurrence of interested outcome for each unit increase of the corresponding predictor, and we use different significance tests, such as likelihood ratio test or Wald test, to examine the significance of predictors. Here we give a brief introduction to the likelihood ratio test.

The likelihood ratio test that we use to assess the model fit as we discussed in section 2.1.2 can also be used to examine the significance of predictors. Let model A be a fitted model that includes the predictor we want to test, and model B is the fitted model that excludes that predictor from model A. Then the test statistic (G) is calculated by subtracting the deviance of model A (D_A) from the deviance of model B (D_B).

$$G = D_A - D_B = -2\log\left(\frac{likelihood\ of\ model\ A}{likelihood\ of\ model\ B}\right) \tag{1.7}$$

There is a significant association between that predictor and the response variable if deviance of model A is significantly smaller. In addition to testing the significance of individual predictor, likelihood ratio test can also be used to test the significance of a set of predictors. The computation of test statistic has the same form as (1.7), and the model A for this test should be a model that includes all the predictors we want to test, and model B excludes those predictors from model A.

2.2 Variable Selection in Logistic Regression

A statistical model is a simplification of reality (Agresti, 2007). At the initial stage of modeling, a large number of candidate predictors are considered to minimize possible modeling biases (Fan and Li, 2006). However, in most cases, not all the predictors have significant effects on the response variable. In statistics, a result is called statistically significant if it is unlikely to have occurred by chance. A simpler model that contains only the important predictors is preferred because it is easy to explain. Parsimony is especially important for high dimension data. The parsimony means that the simplest plausible model with the fewest possible number of predictors is desired.

Variable selection plays an important role in regression analysis and is intended to select the best subset of predictors. There are typically two competing goals in statistical modeling: The model should be complex enough to fit the data well, and also should be simple to interpret (Agresti, 2007). We give a brief summary of three popular variable selection methods for logistic regression: subset selection procedures, information criteria and shrinkage method.

2.2.1 Subset Selection Procedures

The traditional and the most commonly used variable selection methods in logistic regression include backward elimination, forward selection and stepwise selection. Those subset selection procedures produce a subset model at each step by adding or eliminating a predictor from a previous model. Before implementing these procedures, we need to specify a stopping rule: either a cutoff value of the selection criterion or the significance level for the likelihood ratio test or the Wald test for testing the significance of predictors. Each step of these procedures evaluates the current model or the new added predictor and then decides whether we should stop or move on to next step. A final model with the best subset of predictors will be chosen at the end of each procedure.

Backward elimination starts with the full model that includes all candidate predictors. Variables are sequentially deleted from a previous model until the selection criterion of the current model reaches the cutoff value specified in the stopping rule, or all the predictors in the current model are significant. Forward selection, on the contrary, begins with an empty model. The stopping rule for forward selection could be that any added factor would not be significant at a pre-specified significance level. Until the pre-defined stopping rule is satisfied, the most significant predictor of each current model is sequentially added to the model. Stepwise selection modifies the forward selection, where all predictors in the current model are re-evaluated. At each stage of the variable selection process, a predictor might be entered, and another may be eliminated. Also the stepwise selection process ends when it meets a pre-specified stopping rule.

Despite the popularity, there are limitations of these subset selection procedures.

(1) The three different selection procedures can result in different subsets of variables as the “best” model for the same data set.

(2) The correlation among the predictors can result in a final model that is slightly over-fitting, which means that the model is more complex than it should be.

(3) The result is uncertain because only the current model was used to perform statistical inference at each step of the selection while the set of variables included in the current model are very sensitive to the dataset.

(4) The omission of some important predictors can cause bias on the parameter estimation.

(5) Only the nested models, where one model is a subset of another, can be compared.

Therefore, although the subset selection procedures are simple and commonly used in practice, it’s not appropriate to use them in the situations that they give very inconsistent results or when we want to compare the non-nested models, etc.

2.2.2 Information Criteria: AIC and BIC

The Akaike Information Criterion, AIC, (Akaike, 1973) and Bayesian Information Criterion, BIC, (Schwartz, 1978) are the model selection tools based on information theory. The AIC comes from approximately minimizing the difference between the true data distribution and the model distribution, known as the Kullback-Leibler information entropy. While Schwarz derived BIC to asymptotically approximate a transformation of the Bayesian posterior probability of a model. AIC and BIC are given in equation (1.8) and (1.9), respectively, where p is the number of parameters in the model, n is the number of observations, l is the maximum likelihood achieved by the model and $-2\ln(l)$ is the model deviance.

$$AIC = 2 \times p - 2 \times \ln(l) \tag{1.8}$$

$$BIC = p \times \ln(n) - 2 \times \ln(l) \tag{1.9}$$

AIC and BIC both penalize the complexity of the model with an increasing function of number of parameters and also reward the goodness-of-fit. An optimal model achieves the minimum of AIC or BIC. Both information criteria provide a way to compromise between the two competing goals for model building: the model should be complex enough to fit the data adequately, but a simple model is preferred for easy interpretation. However, the penalty term of BIC is more stringent than the penalty term of AIC. Consequently, BIC tends to favor more parsimonious model than AIC does.

When a true model has finite number of candidate predictors and this true model is represented in the list of candidate models, a consistent criterion will asymptotically select the fitted model that has the correct structure with probability one. Whereas if the true model has infinite number of candidate predictors and this true model is not in the list of candidate models, an asymptotically efficient criterion will asymptotically select the fitted model with minimum model deviance. AIC is asymptotically efficient yet not consistent, while BIC is consistent but not asymptotically efficient.

A substantial advantage of AIC and BIC compared with the subset selection procedures is that they can be used to compare non-nested models (Burnham and Anderson, 2002). Also AIC and BIC are able to compare models based on different probability distributions.

2.3 Shrinkage Method

In linear regression, parameter estimation by the ordinary least square (OLS) method is unbiased. However the estimates may have large variance in some cases, the occurrence of multi-collinearity for instance. With slight sacrifice of bias, ridge regression tends to improve the prediction accuracy by shrinking some coefficients. But ridge regression will not shrink values of any coefficients to exact 0, and the fitted model might be too complex to interpret. In 1996, Tibshirani introduced a different shrinkage method, called the Lasso (least absolute shrinkage and selection operator). This method shrinks values of some coefficients to 0 by a constraint on the sum of absolute values of regression coefficients, so Lasso can serve as a tool for variable selection. The substantial difference between Lasso and the subset selection procedures or the information criteria is that Lasso selects variables and estimates the coefficients simultaneously and retains good features of both subset selection and ridge regression.

In Lasso, the constraint on the sum of absolute values regression coefficients is expressed as expression (1.10). p is number of parameters in the model, and t is a positive constant called tuning parameter.

$$Penalty = \sum_{i=1}^{p} |\beta_i| \leq t \tag{1.10}$$

Since the sum of all the coefficients should be less than the value of tuning parameter t, the closer to 0 of t, the more coefficients will shrink towards 0. Therefore, choosing the value of the tuning parameter is crucial for lasso because it controls the model complexity and prediction accuracy.

There are two popular methods to select the optimal value of the tuning parameter that results in a best subset of predictors included in the model. One is to compare the BIC for the resulting models given different value of the tuning parameter. The desired value of the tuning parameter is the one with a minimal BIC. The other is the Cross Validation (CV). The total observations are randomly divided into two parts: training portion and test portion. The training portion is used to fit a model, and then the fitted model is validated by predicting the test portion. The difference between the predicted value and the true value is called cross-validated error. The optimal tuning parameter for a "best" fitted model is the one with minimum mean of cross-validated errors. The objective of this paper is to evaluate and compare these two methods for selecting the optimal tuning parameter through simulation studies.

2.4 L_1 Regularized Logistic Regression

Lasso was originally developed for linear models, and it penalizes the complexity of the model with a constraint on the sum of the absolute values of the model coefficients. For the two-class logistic regression, the lasso modifies the traditional parameter estimation method, maximum log-likelihood, see equation (1.3), with a constraint on the vector of model coefficients as expression (1.10). The lasso solves the following problem:

$$max\left\{ \sum_{i=1}^{n}[y_i(\boldsymbol{\beta}^{\mathrm{T}}x_i) - \log\left(1 + e^{(\boldsymbol{\beta}^{\mathrm{T}}x_i)}\right)]\right\}$$

$$subject\ to\ \sum_{i=1}^{p}|\beta_i| \leq t \qquad (2.1)$$

Another way to state this optimization problem is to add the L_1 norm of the parameter, $\sum_{i=1}^{p}|\beta_i|$, to the objective as following. In this form of L_1 regularization logistic regression, tuning parameter is λ.

$$min\left\{-\sum_{i=1}^{n}\left(y_i(\boldsymbol{\beta}^{\mathrm{T}}x_i)-\log\left(1+e^{(\boldsymbol{\beta}^{\mathrm{T}}x_i)}\right)\right)+\ \lambda\sum_{i=1}^{p}|\beta_i|\right\} \tag{2.2}$$

Un-regularized logistic regression is a convex optimization problem and the objective function is continuously differentiable, so we can use the standard convex optimization methods such as Newton's method to solve it efficiently. The L_1 regularized logistic regression, on the other hand, needs to solve a constrained optimization problem, which is more complex and time consuming. A number of algorithms to solve this optimization problem have been proposed. Generalized lasso was proposed by Roth (2004) and this algorithm develops a generalized lasso algorithm proposed by Osborne (2000). Lee et al. (2006) proposed an efficient algorithm that interactively approximates the objective function by a quadratic approximation at the current point, and maintains the L_1 constraint at the same time.

More recently, Friedman, Hastie and Tibshirani (2009) developed an efficient algorithm called coordinate descent for estimation of generalized linear model with convex penalties. Also they provide an R package $glmnet$ which is publicly available. We used this R package $glmnet$ to estimate the coefficients with the coordinate descent algorithm in all simulation studies.

As the values of tuning parameter increases, more coefficients will shrink to zero. For example, we simulated a dataset with 200 observations based on the model that has a binary response y, eight normally distributed predictors with mean 0 and standard deviation 1, the true model is

$$\ln\left(\frac{\mathrm{P}(y=1)}{1-\mathrm{P}(\mathrm{y}=1)}\right)=x_1+x_2+x_3+x_4+0x_5+0x_6+0x_7+0x_8$$

Then, we used the glmnet function available in the $glmnet$ package to fit a series of L_1 regularized logistic regression models with the values of λ suggested by $glmnet$. Figure 2 is the

plot of coefficient estimates against the values of the tuning parameter λ. The scale at the top is the number of predictors left in the model.

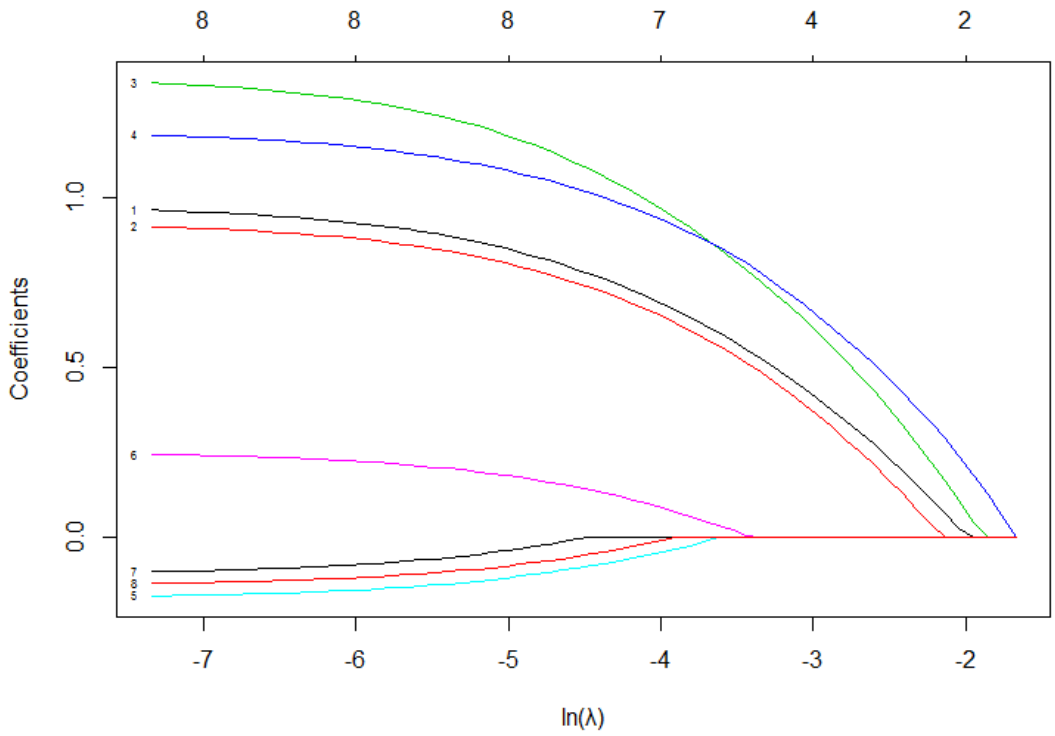

Figure 2 Relationship of the Coefficients Estimates and $\ln(\lambda)$ in L_1 regularized logistic regression

When $\ln(\lambda) = -4$, $\lambda = 0.0183$, coefficient of x_8 turns into 0, seven predictors left in the model. When $\ln(\lambda) = -3.4$, $\lambda = 0.0334$, coefficients of x_5, x_6, x_7 and x_8 turn into 0, and predictors x_1, x_2, x_3 and x_4 are left in the model.

2.5 Selection of Tuning Parameter

2.5.1 Cross Validation

Cross validation is a popular method for estimating the prediction error and comparing different models. Typically, we would partition the dataset into two parts: the training data and the testing data. In k-fold cross validation, the dataset will be randomly split into k mutually exclusive subsets of approximately equal size.

Among the k subsets, one subset is retained as validation data for testing the model, and the remaining k-1 subsets are used as training data to fit the model. The cross validation process is repeated k times, and each of the subsets is used exactly once as validation data. Different values of the tuning parameter could result in different fitted model using the same training data. The optimal model is the one that has the minimum cross-validated errors, and the corresponding value of the tuning parameter for the optimal model is preferred.

2.5.2 Bayesian Information Criterion

For L_1 regularized logistic regression model, we estimate the coefficients using penalized maximum likelihood estimation (2.2) given the value of the tuning parameter. Bayesian Information Criterion (BIC) compares models based on the deviance. BIC of a logistic regression model is calculated as follows:

$$BIC = p \times \ln(n) - 2\ln(l) = p \times \ln(n) + deviance \qquad (2.3)$$

BIC penalizes the model complexity with term $p \times \ln(n)$. For two models with same deviance, the model that includes less number of parameters has smaller BIC value. The variable

selection process based on BIC compares the models resulted in different tuning parameter values. The optimal fitted model is identified by the minimum value of BIC.

2.6 Simulation Study

The purpose of this simulation is to examine the performance of the CV and BIC for choosing the optimal value of tuning parameter in terms of variable selection and also prediction accuracy. Their performance in terms of variable selection is evaluated by the True Positive Rate (TPR) and False Positive Rate (FPR). TPR is the proportion of correctly selected important predictors among the true important predictors, while the FPR is the proportion of falsely selected important predictors among the true unimportant predictors. Both TPR and FPR have value ranging from 0 to 1. A model that has both TPR closer to 1 and FPR closer to 0 is desired.

Their performance in terms of prediction accuracy is evaluated by the sum of absolute difference between the estimated coefficients and true coefficients, which is called bias for the sake of simplicity. A model with smaller bias estimates the coefficients more accurately. However, since a model that has more true important predictors tends to have larger bias, this estimation accuracy evaluation criterion is adjusted by dividing the bias by the number of important predictors in the true model, called adjusted bias.

2.6.1 The Model Setup

To compare the performance of CV and BIC for selecting the optimal tuning parameter, we generated the data consisting n observations based on the models with different number of predictors (p), proportion of important predictors among all predictors (p_1/p). The $AR(1)$ correlation structure with different correlation coefficient (ρ) was used.

In practice, the important predictors, although they all significantly affect the response, may have different level of influence on the response variable, so we set up two different structures for the vector of the coefficients ($\boldsymbol{\beta}$).

2.6.2 Design of the Simulation

We chose the models with number of predictors (p), proportion of important predictors among all predictors (p_1/p), number of observations (n), correlation coefficient (ρ) and the vector of coefficients ($\boldsymbol{\beta}$) as follow.

1) $p = 10\ or\ 20$
2) $p_1/p = 0.8\ or\ 0.2$
3) $n = 100\ or\ 200\ or\ 400$
4) $\rho = 0\ or\ 0.5\ or\ 0.8\ or\ 0.95\ or\ 0.99$

When ρ=0, the predictors are independent and identically distributed (IID).

5) $\boldsymbol{\beta_1} = (\underbrace{1, \ldots, 1}_{p_1}, \underbrace{0, \ldots, 0}_{p-p_1})$ or $\boldsymbol{\beta_2} = (\underbrace{1, \ldots, 1}_{p_1/2}, \underbrace{0.5, \ldots, 0.5}_{p_1/2}, \underbrace{0, \ldots, 0}_{p-p_1})$

We call the vector of coefficients as $\boldsymbol{\beta_1}$ if the first p_1 elements are 1 and the last $p - p_1$ elements are 0. While the vector of coefficients is called as $\boldsymbol{\beta_2}$ if the first $p_1/2$ elements are 1, the following $p_1/2$ elements are 0.5 and the last $p - p_1$elements are 0. For instance, a model has 4 important predictors out of total 6 predictors and the structure of the coefficients is $\boldsymbol{\beta_1}$, then the logit of this model is

$$logit = x_1 + x_2 + x_3 + x_4 + 0x_5 + 0x_6$$

If the structure of the coefficients of this model is $\boldsymbol{\beta_2}$, then the logit of this model is

$$logit = x_1 + x_2 + 0.5x_3 + 0.5x_4 + 0x_5 + 0x_6$$

In addition, we are interested in their performance on the complex data that has a large number of predictors, called high dimension data. So we also simulated data with 400 or 800 observations based on the model with 50 predictors, and 20% of these predictors are important.

Under each scenario, we simulated 200 data sets. All simulations were conducted using the $glmnet$ package available in R 2.15.1, which is developed by Friedman, Hastie and Tibshirani. This package provides extremely efficient procedures for fitting the lasso regularization path for logistic regression models.

2.6.3 The Procedure

The following procedures explain how the BIC and CV methods select the value for tuning parameter λ and how we compute the bias, adjusted bias, TPR and FPR with the coefficients estimates of the selected model.

1) Simulate one dataset consisting n observations based on the model with p_1 important predictors among p candidate predictors and the structure of the vector of coefficients is $\boldsymbol{\beta_1}$ or $\boldsymbol{\beta_2}$. All predictors are normally distributed with mean 0 and standard deviation 1, and they are correlated in the $AR(1)$ correlation structure with correlation coefficient ρ.

2) With the simulated data and the value of tuning parameter λ suggested by the $glmnet$ function in $glmnet$ R package, fit a penalized L_1 logistic model and get the coefficients estimates and model deviance.

3) Compute the BIC for each subset model by adding the model deviance and multiplication of p and $log(n)$, then find the subset model with the smallest BIC and its estimated coefficients. Then compute the bias, adjusted bias, TPR and FPR of this selected model.

4) The $glmnet$ package provides the $cv.glmnet$ function to perform a 10-fold cross validation and returns an optimal value for the tuning parameter λ, which results in a model with minimum mean cross validated error (cvm). With the same simulated data, we use this function and get the optimal value of tuning parameter λ and its corresponding coefficients estimation. Then compute the bias, adjusted bias, TPR and FPR of this selected model;

5) Repeat 1) through 4) for 200 times and at each time the dataset generated has the same scenario with the dataset generated in 1). Then calculate the mean and standard deviation of the bias, adjusted bias, TPR and FPR for each method.

CHAPTER 3

Result and Conclusion

3.1 Simulation Result

We conducted the simulations and summarized the results in seven sets. In simulation set A, we kept the structure of the vector of coefficients vector as $\boldsymbol{\beta_1}$ and the predictors are independent and identically distributed (IID). Table 2 is the summary result of this simulation set. When $p_1/p = 0.8$, $p = 20$ and $n = 100$, some fitted models did not converge. The possible reason is that the ratio of the number of predictors and sample size is high. We mark all the non-convergence cases with N/A, and we will discuss the non-convergence issues in detail in section 3.3.

Except the non-convergence, Table 2 shows very consistent results. BIC gives smaller bias, higher TPR and smaller standard deviation of the TPR; while CV gives smaller standard deviation of the bias, lower FPR and smaller standard deviation of the FPR. Therefore, BIC outperforms CV regarding the bias, TPR and standard deviation of the TPR, while CV performs better regarding the standard deviation of the bias, FPR and standard deviation of the FPR. The only exception here is that when $p_1/p = 0.8$, $p = 10$ and $n = 400$, BIC has a smaller standard deviation of the bias.

Table 2 Result of Simulation Set A with Model Setup as $\boldsymbol{\beta_1}$ and $\rho = 0$

n	p	p_1/p	Method	Bias	Standard Deviation of Bias	Adjusted Bias	Standard Deviation of Adjusted Bias	TPR	Standard Deviation of TPR	FPR	Standard Deviation of FPR
100	10	0.8	BIC	3.7903	2.5862	0.4738	0.3233	0.9919	0.0378	0.6050	0.3565
			CV	4.8796	0.9017	0.6100	0.1127	0.9481	0.0981	0.2325	0.3320
		0.2	BIC	1.0216	0.4959	0.5108	0.2480	0.9550	0.1750	0.0881	0.1307
			CV	1.3015	0.3124	0.6508	0.1562	0.9175	0.2285	0.0331	0.0757
	20	0.8	BIC	N/A							
			CV								
		0.2	BIC	1.0181	0.4400	0.5091	0.2200	0.9650	0.1546	0.1081	0.1399
			CV	1.2961	0.2966	0.6481	0.1483	0.9525	0.1707	0.0444	0.0821
200	10	0.8	BIC	2.1580	0.8294	0.2697	0.1037	1.0000	0.0000	0.6475	0.3573
			CV	3.9721	0.7201	0.4965	0.0900	1.0000	0.0000	0.2175	0.2772
		0.2	BIC	0.7337	0.2626	0.3669	0.1313	1.0000	0.0000	0.0813	0.1119
			CV	1.0528	0.2343	0.5264	0.1172	0.9975	0.0354	0.0206	0.0557
	20	0.8	BIC	6.6026	4.0100	0.4127	0.2506	0.9988	0.0088	0.7750	0.2351
			CV	8.9240	1.1806	0.5578	0.0738	0.9950	0.0181	0.3188	0.2282
		0.2	BIC	1.8739	0.4421	0.4685	0.1105	1.0000	0.0000	0.1013	0.1006
			CV	2.1536	0.3719	0.5384	0.0930	0.9988	0.0177	0.0653	0.0760
400	10	0.8	BIC	1.4741	0.5110	0.1843	0.0639	1.0000	0.0000	0.6300	0.3481
			CV	3.2117	0.6160	0.4015	0.0770	1.0000	0.0000	0.2100	0.2938
		0.2	BIC	0.5510	0.1892	0.2755	0.0946	1.0000	0.0000	0.0688	0.1016
			CV	0.8651	0.1744	0.4325	0.0872	1.0000	0.0000	0.0169	0.0480
	20	0.8	BIC	3.6373	1.4316	0.2273	0.0895	1.0000	0.0000	0.7563	0.2549
			CV	7.1566	1.1119	0.4473	0.0695	1.0000	0.0000	0.3688	0.2480
		0.2	BIC	1.4401	0.3272	0.3600	0.0818	1.0000	0.0000	0.0809	0.0865
			CV	1.7669	0.3027	0.4417	0.0757	1.0000	0.0000	0.0450	0.0726

Table 3 is the summary result for simulation set B, in which we kept the model coefficients vector as $\boldsymbol{\beta_2}$ and the predictors are independent and identically distributed (IID). As the result of simulation set B, BIC is better in terms of bias, TPR and standard deviation of TPR, while CV is better in terms of standard deviation of the bias, FPR and standard deviation of FPR. However, same as simulation set A, when $p_1/p = 0.8$, $p = 10$ and $n = 400$, BIC has a smaller standard deviation of the bias. Besides, when $p_1/p = 0.2$, $p = 20$ and $n = 100$, BIC has lower

TPR and larger standard deviation of TPR, while CV has higher FPR and larger standard deviation of FPR.

Table 3 Result of Simulation Set B with Model Setup as $\boldsymbol{\beta_2}$ and $\rho = 0$

n	p	p_1/p	Method	Bias	Standard Deviation of Bias	Adjusted Bias	Standard Deviation of Adjusted Bias	TPR	Standard Deviation of TPR	FPR	Standard Deviation of FPR
100	10	0.8	BIC	3.1454	1.2296	0.3932	0.1537	0.8563	0.1957	0.4575	0.3908
			CV	3.9936	0.7785	0.4992	0.0973	0.7513	0.1969	0.1700	0.2811
		0.2	BIC	0.9390	0.4221	0.4695	0.2111	0.7125	0.3048	0.0763	0.1325
			CV	1.0900	0.2666	0.5450	0.1333	0.6900	0.3104	0.0513	0.1101
	20	0.8	BIC	N/A							
			CV								
		0.2	BIC	2.0437	0.5132	0.5109	0.1283	0.6525	0.2855	0.0566	0.0809
			CV	2.2183	0.3898	0.5546	0.0974	0.6838	0.2384	0.0597	0.0837
200	10	0.8	BIC	1.7906	0.6629	0.2238	0.0829	0.9831	0.0597	0.4625	0.4039
			CV	3.0989	0.5775	0.3874	0.0722	0.9188	0.0969	0.1350	0.2441
		0.2	BIC	0.6546	0.3163	0.3273	0.1581	0.9250	0.1790	0.0763	0.1194
			CV	0.8902	0.2160	0.4451	0.1080	0.8525	0.2340	0.0388	0.0924
	20	0.8	BIC	5.1437	1.9938	0.3215	0.1246	0.9628	0.0750	0.6325	0.3099
			CV	6.7466	1.0806	0.4217	0.0675	0.9028	0.0935	0.2825	0.2596
		0.2	BIC	1.5752	0.4002	0.3938	0.1000	0.8775	0.1754	0.0803	0.0949
			CV	1.7813	0.3377	0.4453	0.0844	0.8600	0.1818	0.0597	0.0837
400	10	0.8	BIC	1.2509	0.3562	0.1564	0.0445	1.0000	0.0000	0.5625	0.3913
			CV	2.5564	0.4661	0.3196	0.0583	0.9894	0.0371	0.1800	0.2747
		0.2	BIC	0.4744	0.1740	0.2372	0.0870	0.9900	0.0702	0.0594	0.0929
			CV	0.7337	0.1689	0.3669	0.0845	0.9400	0.1629	0.0131	0.0404
	20	0.8	BIC	3.0542	0.9174	0.1909	0.0573	0.9978	0.0115	0.6825	0.2776
			CV	5.4874	0.8716	0.3430	0.0545	0.9838	0.0302	0.2913	0.2478
		0.2	BIC	1.1675	0.2836	0.2919	0.0709	0.9925	0.0428	0.0906	0.0879
			CV	1.4549	0.2504	0.3637	0.0626	0.9688	0.0866	0.0425	0.0719

Correlation coefficient ρ was set as 0.5 for simulation set C and D. Simulation set C has $\boldsymbol{\beta_1}$ and simulation set D has $\boldsymbol{\beta_2}$. The results displayed in Table 4 and 5 indicate that BIC is

better in terms of bias, TPR and standard deviation of TPR, while CV is better in terms of standard deviation of the bias, FPR and standard deviation of FPR.

Table 4 Result of Simulation Set C with Model Setup as $\boldsymbol{\beta_1}$ and $\rho = 0.5$

n	p	p_1/p	Method	Bias	Standard Deviation of Bias	Adjusted Bias	Standard Deviation of Adjusted Bias	TPR	Standard Deviation of TPR	FPR	Standard Deviation of FPR
100	10	0.8	BIC	N/A							
			CV								
		0.2	BIC	0.9415	0.6937	0.4708	0.3468	0.9975	0.0354	0.0975	0.1402
			CV	1.1534	0.2609	0.5767	0.1304	0.9800	0.0982	0.0344	0.0674
	20	0.8	BIC	N/A							
			CV								
		0.2	BIC								
			CV								
200	10	0.8	BIC	3.0178	1.2486	0.3772	0.1561	0.9944	0.0288	0.5325	0.3787
			CV	3.9681	0.6680	0.4960	0.0835	0.9869	0.0423	0.1425	0.2623
		0.2	BIC	0.6941	0.2551	0.3471	0.1276	1.0000	0.0000	0.0700	0.1075
			CV	1.0301	0.1948	0.5151	0.0974	0.9950	0.0499	0.0094	0.0414
	20	0.8	BIC	N/A							
			CV								
		0.2	BIC	1.6497	0.3953	0.4124	0.0988	0.9950	0.0351	0.0869	0.0895
			CV	1.9986	0.3585	0.4996	0.0896	0.9900	0.0491	0.0447	0.0702
400	10	0.8	BIC	1.9683	0.5612	0.2460	0.0702	1.0000	0.0000	0.4850	0.3646
			CV	3.3590	0.5508	0.4199	0.0689	0.9994	0.0088	0.1350	0.2336
		0.2	BIC	0.4951	0.1899	0.2475	0.0949	1.0000	0.0000	0.0650	0.0946
			CV	0.8232	0.1685	0.4116	0.0843	1.0000	0.0000	0.0125	0.0434
	20	0.8	BIC	5.4143	2.3651	0.3384	0.1478	0.9994	0.0062	0.6363	0.2844
			CV	7.6377	1.0174	0.4774	0.0636	0.9984	0.0116	0.2288	0.2239
		0.2	BIC	1.2765	0.3031	0.3191	0.0758	1.0000	0.0000	0.0766	0.0804
			CV	1.6639	0.2799	0.4160	0.0700	1.0000	0.0000	0.0272	0.0500

Table 5 Result of Simulation Set D with Model Setup as $\boldsymbol{\beta_2}$ and $\rho = 0.5$

n	p	p_1/p	Method	Bias	Standard Deviation of Bias	Adjusted Bias	Standard Deviation of Adjusted Bias	TPR	Standard Deviation of TPR	FPR	Standard Deviation of FPR
100	10	0.8	BIC	N/A							
			CV								
		0.2	BIC	0.8206	0.4347	0.4103	0.2174	0.8275	0.2435	0.0756	0.1234
			CV	0.9703	0.2105	0.4852	0.1052	0.7550	0.2604	0.0319	0.0841
	20	0.8	BIC	N/A							
			CV								
		0.2	BIC								
			CV								
200	10	0.8	BIC	2.4886	1.1343	0.3111	0.1418	0.9538	0.0734	0.4300	0.3685
			CV	3.0479	0.5302	0.3810	0.0663	0.9144	0.0900	0.1200	0.2255
		0.2	BIC	0.5839	0.2461	0.2920	0.1230	0.9450	0.1568	0.0744	0.1100
			CV	0.8088	0.1763	0.4044	0.0882	0.8825	0.2125	0.0238	0.0593
	20	0.8	BIC	N/A							
			CV								
		0.2	BIC	1.2533	0.3213	0.3133	0.0803	0.9300	0.1206	0.0519	0.0728
			CV	1.5441	0.2925	0.3860	0.0731	0.8938	0.1449	0.0184	0.0508
400	10	0.8	BIC	1.6001	0.5269	0.2000	0.0659	0.9944	0.0260	0.4475	0.3697
			CV	2.5235	0.4206	0.3154	0.0526	0.9769	0.0487	0.1075	0.2119
		0.2	BIC	0.3988	0.1613	0.1994	0.0807	0.9900	0.0702	0.0644	0.1002
			CV	0.6557	0.1328	0.3278	0.0664	0.9725	0.1143	0.0075	0.0346
	20	0.8	BIC	4.1588	1.2247	0.2599	0.0765	0.9797	0.0343	0.5613	0.2928
			CV	5.5960	0.7269	0.3498	0.0454	0.9656	0.0455	0.1838	0.1917
		0.2	BIC	0.9856	0.2500	0.2464	0.0625	0.9888	0.0520	0.0538	0.0712
			CV	1.3098	0.2245	0.3274	0.0561	0.9700	0.0852	0.0178	0.0511

For simulation set E and F, correlation coefficient ρ was set as 0.8. Simulation set C has $\boldsymbol{\beta_1}$ and simulation set D has $\boldsymbol{\beta_2}$. Table 6 and Table 7 display the results for simulation set E and F, which seem that they have the same overall results as simulation set A – D, BIC performs better in terms of bias, TPR and standard deviation of TPR, while CV performs better in terms of

standard deviation of the bias, FPR and standard deviation of FPR. Exception here is that for models with $p_1/n = 8/200 = 16/400$, CV gets smaller bias that BIC.

Table 6 Result of Simulation Set E with Model Setup as $\boldsymbol{\beta_1}$ and $\rho = 0.8$

n	p	p_1/p	Method	Bias	Standard Deviation of Bias	Adjusted Bias	Standard Deviation of Adjusted Bias	TPR	Standard Deviation of TPR	FPR	Standard Deviation of FPR
100	10	0.8	BIC	N/A							
			CV								
		0.2	BIC	1.2049	0.7732	0.6025	0.3866	0.9325	0.1713	0.1225	0.1575
			CV	1.2505	0.3155	0.6253	0.1578	0.9025	0.1986	0.0450	0.0763
	20	0.8	BIC	N/A							
			CV								
		0.2	BIC								
			CV								
200	10	0.8	BIC	5.1295	4.7349	0.6412	0.5919	0.9188	0.0884	0.3925	0.3320
			CV	4.3571	0.5607	0.5446	0.0701	0.8950	0.0958	0.1725	0.2681
		0.2	BIC	0.8002	0.3819	0.4001	0.1910	0.9925	0.0609	0.0925	0.1164
			CV	1.0287	0.2238	0.5144	0.1119	0.9875	0.0783	0.0363	0.0737
	20	0.8	BIC	N/A							
			CV								
		0.2	BIC	1.8012	0.5062	0.4503	0.1266	0.9538	0.1005	0.0703	0.0792
			CV	2.0599	0.3668	0.5150	0.0917	0.9425	0.1113	0.0309	0.0509
400	10	0.8	BIC	3.1205	1.1793	0.3901	0.1474	0.9838	0.0440	0.3700	0.3334
			CV	3.6516	0.5827	0.4565	0.0728	0.9756	0.0570	0.2025	0.2840
		0.2	BIC	0.5626	0.2488	0.2813	0.1244	1.0000	0.0000	0.0900	0.1183
			CV	0.8457	0.1619	0.4228	0.0810	1.0000	0.0000	0.0213	0.0487
	20	0.8	BIC	13.1369	26.8904	0.8211	1.6807	0.9400	0.0567	0.5163	0.3171
			CV	8.5609	1.1904	0.5351	0.0744	0.9338	0.0606	0.1913	0.2112
		0.2	BIC	1.3493	0.3740	0.3373	0.0935	0.9963	0.0305	0.0794	0.0751
			CV	1.7161	0.2972	0.4290	0.0743	0.9950	0.0351	0.0263	0.0479

Table 7 Result of Simulation Set F with Model Setup as $\boldsymbol{\beta_2}$ and $\rho = 0.8$

n	p	p_1/p	Method	Bias	Standard Deviation of Bias	Adjusted Bias	Standard Deviation of Adjusted Bias	TPR	Standard Deviation of TPR	FPR	Standard Deviation of FPR
100	10	0.8	BIC	N/A							
			CV								
		0.2	BIC	0.8388	0.4525	0.4194	0.2262	0.8325	0.2366	0.0775	0.1030
			CV	0.9768	0.2423	0.4884	0.1212	0.7625	0.2553	0.0363	0.0737
	20	0.8	BIC	N/A							
			CV								
		0.2	BIC								
			CV								
200	10	0.8	BIC	3.4683	1.2562	0.4335	0.1570	0.8513	0.1059	0.3275	0.3530
			CV	3.3085	0.4808	0.4136	0.0601	0.8213	0.1082	0.1475	0.2496
		0.2	BIC	0.6487	0.3135	0.3244	0.1568	0.9325	0.1713	0.0794	0.1136
			CV	0.8035	0.1867	0.4018	0.0934	0.8775	0.2156	0.0269	0.0662
	20	0.8	BIC	N/A							
			CV								
		0.2	BIC	1.4632	0.4400	0.3658	0.1100	0.8788	0.1349	0.0625	0.0698
			CV	1.5927	0.2660	0.3982	0.0665	0.8525	0.1487	0.0244	0.0442
400	10	0.8	BIC	2.4893	0.7680	0.3112	0.0960	0.9344	0.0792	0.3150	0.3337
			CV	2.8171	0.4017	0.3521	0.0502	0.9069	0.0938	0.1275	0.2605
		0.2	BIC	0.4870	0.2110	0.2435	0.1055	0.9850	0.0855	0.0744	0.0851
			CV	0.6606	0.1435	0.3303	0.0717	0.9500	0.1504	0.0244	0.0497
	20	0.8	BIC	6.3251	1.8855	0.3953	0.1178	0.8856	0.0706	0.3413	0.2631
			CV	6.2647	0.7322	0.3915	0.0458	0.8672	0.0725	0.1475	0.1862
		0.2	BIC	1.0367	0.3360	0.2592	0.0840	0.9600	0.0919	0.0569	0.0547
			CV	1.3137	0.2193	0.3284	0.0548	0.9313	0.1119	0.0144	0.0318

To simulate the cases that multi-collinearity exists among the predictors, correlation coefficient ρ was set as 0.95 for simulation set G and as 0.99 for simulation set H. Since non-convergence occurs for most cases where $p_1/p = 0.8$, we only considered the cases that $p_1/p = 0.2$. The results displayed in Table 8 and Table 9 indicate that overall BIC is better in terms

of TPR and standard deviation of TPR, while CV is better in terms of bias, standard deviation of the bias, FPR and standard deviation of FPR.

Table 8 Result of Simulation Set G with Model Setup as $p_1/p = 0.2$ and $\rho = 0.95$

$\boldsymbol{\beta}$	n	p	Method	Bias	Standard Deviation of Bias	Adjusted Bias	Standard Deviation of Adjusted Bias	TPR	Standard Deviation of TPR	FPR	Standard Deviation of FPR
$\boldsymbol{\beta_1}$	100	10	BIC	1.8225	1.1823	0.9113	0.5911	0.7150	0.2629	0.1331	0.1373
			CV	1.4723	0.4519	0.7361	0.2260	0.6750	0.2688	0.0775	0.1007
		20	BIC	4.3403	7.0126	1.0851	1.7532	0.6638	0.1729	0.0931	0.0917
			CV	2.9292	1.4622	0.7323	0.3655	0.6500	0.1807	0.0506	0.0750
	200	10	BIC	1.2479	0.6690	0.6240	0.3345	0.8725	0.2185	0.1181	0.1136
			CV	1.2049	0.3629	0.6024	0.1814	0.8350	0.2357	0.0706	0.0891
		20	BIC	2.7406	1.1243	0.6852	0.2811	0.8138	0.1662	0.0856	0.0760
			CV	2.4373	0.5340	0.6093	0.1335	0.7875	0.1732	0.0506	0.0622
	400	10	BIC	0.9624	0.5620	0.4812	0.2810	0.9700	0.1190	0.1263	0.1215
			CV	0.9977	0.2859	0.4989	0.1429	0.9575	0.1398	0.0731	0.0915
		20	BIC	2.2933	0.9125	0.5733	0.2281	0.8988	0.1375	0.0756	0.0685
			CV	2.1578	0.5074	0.5395	0.1268	0.8875	0.1499	0.0397	0.0490
$\boldsymbol{\beta_2}$	100	10	BIC	1.5140	2.4719	0.7570	1.2360	0.6575	0.2632	0.1156	0.1308
			CV	1.1571	0.4656	0.5785	0.2328	0.6175	0.2698	0.0606	0.0938
		20	BIC	N/A							
			CV								
	200	10	BIC	0.9937	0.5146	0.4969	0.2573	0.7850	0.2482	0.1006	0.1105
			CV	0.9673	0.2999	0.4837	0.1500	0.7525	0.2506	0.0531	0.0767
		20	BIC	2.1994	0.8786	0.5499	0.2196	0.7063	0.1726	0.0706	0.0693
			CV	1.9331	0.4379	0.4833	0.1095	0.6688	0.1789	0.0331	0.0451
	400	10	BIC	0.7943	0.3972	0.3971	0.1986	0.8250	0.2391	0.1038	0.1108
			CV	0.8008	0.2436	0.4004	0.1218	0.7925	0.2470	0.0563	0.0866
		20	BIC	1.7332	0.6161	0.4333	0.1540	0.8100	0.1529	0.0759	0.0660
			CV	1.6027	0.3785	0.4007	0.0946	0.7963	0.1526	0.0347	0.0468

Table 9 Result of Simulation Set H with Model Setup as $p_1/p = 0.2$ and $\rho = 0.99$

$\boldsymbol{\beta}$	n	p	Method	Bias	Standard Deviation of Bias	Adjusted Bias	Standard Deviation of Adjusted Bias	TPR	Standard Deviation of FPR	FPR	Standard Deviation of FPR
$\boldsymbol{\beta_1}$	100	10	BIC	2.4868	1.8741	1.2434	0.9371	0.5500	0.2745	0.1281	0.1178
			CV	1.7782	0.9675	0.8891	0.4837	0.5075	0.2675	0.0981	0.1032
		20	BIC	5.3582	3.2656	1.3396	0.8164	0.4450	0.1647	0.0853	0.0752
			CV	3.4747	0.8052	0.8687	0.2013	0.4500	0.1825	0.0631	0.0645
	200	10	BIC	2.0443	1.0062	1.0221	0.5031	0.6250	0.2641	0.1394	0.1216
			CV	1.5943	0.6418	0.7971	0.3209	0.5850	0.2658	0.1063	0.1083
		20	BIC	4.3439	1.4143	1.0860	0.3536	0.5463	0.1943	0.0850	0.0672
			CV	3.2312	0.7921	0.8078	0.1980	0.5400	0.2012	0.0703	0.0657
	400	10	BIC	1.7044	0.8148	0.8522	0.4074	0.7525	0.2556	0.1519	0.1252
			CV	1.3831	0.5395	0.6915	0.2698	0.7325	0.2599	0.1069	0.1066
		20	BIC	3.6605	1.2499	0.9151	0.3125	0.6588	0.1912	0.0884	0.0684
			CV	2.9465	0.8151	0.7366	0.2038	0.6488	0.1927	0.0681	0.0661
$\boldsymbol{\beta_2}$	100	10	BIC	2.0813	1.7247	1.0406	0.8623	0.4600	0.2574	0.1225	0.1141
			CV	1.4137	0.5226	0.7069	0.2613	0.4100	0.2593	0.0881	0.0961
		20	BIC	3.9380	1.4646	0.9845	0.3662	0.3688	0.1584	0.0741	0.0636
			CV	2.7300	0.6769	0.6825	0.1692	0.3563	0.1554	0.0566	0.0534
	200	10	BIC	1.7622	1.2395	0.8811	0.6198	0.5525	0.2476	0.1325	0.1213
			CV	1.3004	0.5531	0.6502	0.2765	0.5125	0.2478	0.1006	0.1097
		20	BIC	3.4510	1.4918	0.8627	0.3730	0.4950	0.1973	0.0766	0.0690
			CV	2.5374	0.7063	0.6343	0.1766	0.4738	0.1932	0.0547	0.0588
	400	10	BIC	1.3118	0.8059	0.6559	0.4029	0.6400	0.2707	0.1319	0.1229
			CV	1.0977	0.5099	0.5489	0.2549	0.6225	0.2676	0.0950	0.1043
		20	BIC	2.8258	0.9989	0.7065	0.2497	0.5913	0.1928	0.0747	0.0629
			CV	2.2592	0.6805	0.5648	0.1701	0.5613	0.1850	0.0559	0.0609

Last, in the high dimension case, each data has 50 predictors. Simulation result shown in Table 10 still indicates that BIC performs better in terms of bias, TPR and standard deviation of TPR, while CV performs better in terms of standard deviation of the bias, FPR and standard deviation of FPR. However, BIC now outperforms CV regarding the FPR and standard deviation of FPR when $\rho = 0.5$ and $n = 400$ or $\rho = 0$.

Table 10 Result of High Dimension Case with Model Setup as $p_1/p = 0.2$ and $p = 50$

p	p_1/p	ρ	n	Method	Bias	Standard Deviation of Bias	Adjusted Bias	Standard Deviation of Adjusted Bias	TPR	Standard Deviation of TPR	FPR	Standard Deviation of FPR
50	0.2	0	400	BIC	5.2689	0.6457	0.5269	0.0646	1.0000	0.0000	0.1088	0.0769
				CV	5.2725	0.5282	0.5273	0.0528	1.0000	0.0000	0.1324	0.0799
			800	BIC	4.2315	0.5354	0.4231	0.0535	1.0000	0.0000	0.1106	0.0749
				CV	4.3045	0.4328	0.4305	0.0433	1.0000	0.0000	0.1209	0.0830
		0.5	400	BIC	4.8942	0.5945	0.4894	0.0594	1.0000	0.0000	0.0740	0.0595
				CV	5.0262	0.5524	0.5026	0.0552	1.0000	0.0000	0.0841	0.0662
			800	BIC	4.0391	0.5300	0.4039	0.0530	1.0000	0.0000	0.0696	0.0601
				CV	4.2414	0.4626	0.4241	0.0463	1.0000	0.0000	0.0621	0.0551
		0.8	400	BIC	4.8043	0.6009	0.4804	0.0601	0.9710	0.0497	0.0655	0.0516
				CV	5.1444	0.5499	0.5144	0.0550	0.9685	0.0507	0.0545	0.0491
			800	BIC	3.9678	0.5452	0.3968	0.0545	0.9980	0.0140	0.0778	0.0512
				CV	4.3667	0.5256	0.4367	0.0526	0.9975	0.0157	0.0558	0.0469

3.2 Conclusion

By varying the number of observations (n), number of predictors (p), proportion of important predictors among all predictors (p_1/p), the strength of correlation between the predictors (ρ) and the structure of the vector of coefficients ($\boldsymbol{\beta}$), we compared the performances of BIC and CV on choosing the tuning parameter of the L_1 regularized logistic regression in terms of parameter estimation and variable selection. Overall, simulation results show that:

1) BIC achieves smaller bias, higher TPR and smaller standard deviation of TPR.

2) CV achieves smaller standard deviation of bias, lower FPR and smaller standard deviation of FPR.

However, when serious multi-collinearity exists among the predictors, the CV performs better with smaller bias. The simulation result indicates that BIC achieves better prediction accuracy, and it has more consistency and power to choose the true important predictors, while CV achieves better estimation accuracy when serious multi-collinearity exists, and it tends to get more consistent coefficients estimation, although not as accurate as BIC, and also performs better screening out the false important predictors correctly and consistently.

However there are several exceptions:

1) BIC outperforms CV with smaller standard deviation of the bias when p_1=8, n=400 and predictors are IID.

2) BIC outperforms CV with lower FPR and smaller standard deviation of the FPR when $p_1 = 4, n = 100$, predictors are IID and true model has coefficients vector $\boldsymbol{\beta_2}$.

3) CV outperforms BIC with higher TPR and smaller standard deviation of the TPR when $p_1 = 4, n = 100$, predictors are IID and true model has coefficients vector $\boldsymbol{\beta_2}$.

4) CV outperforms BIC with smaller bias when $p_1/n = 8/200 = 16/400$ and predictors have a correlation coefficient 0.8.

We also find that a high ratio of the number of predictors and number of observations and/or existence of multi-collinearity could cause non-convergence of the fitted models.

3.3 Remarks on the Non-convergence Issue

According to the simulation result, some fitted models didn't reach convergence. As we mentioned in section 2.2.1, the possible reasons for non-convergence include but are not limited

to the high ratio of the number of predictors and sample size, multi-collinearity and separation problem. Next we use two non-convergence cases as examples, analyze the reason that causes the non-convergence and the possible solutions.

3.3.1 Example 1

When the models are setup as $\rho = 0, p = 20, p1/p = 0.8$ and $n = 100$, the coefficients estimation of some fitted models didn't reach convergence. Here multi-collinearity is not the reason for non-convergence because all the predictors are independent and identically distributed. Using one dataset with non-converged fitted L_1 regularized logistic regression model, we conducted a linear discriminate analysis to find the linear combination of the predictors which separates the two classes of the response variable. This classification procedure results in a misclassification rate 5%, which means that it didn't completely separate two classes of the response variable. Thus, separation is not the reason for non-convergence.

Next, using the same dataset, we fitted a logistic regression model without the L_1 penalty, the maximum likelihood estimators didn't converge neither. However, with a larger sample size, 200, all the fitted models converged. So the reason that causes the non-convergence of this case is high ratio of number of predictors and sample size. The solution for the non-convergence caused by this reason can be solved with an increased sample size. In general, the logistic regression models require approximately 10 observations per predictor to reach convergence (Peduzzi, 1996).

3.3.2 Example 2

When the models are setup as $\rho = 0.8, p = 10, p1/p = 0.8$ and $n = 100$, the coefficients estimation of some fitted models didn't reach convergence. Since when the models are setup as $\rho = 0, p = 10, p1/p = 0.8$ and $n = 100$, all the fitted models converged, collinearity might be the reason for non-convergence when predictors are correlated. We checked whether the non-convergence problem can be solved if we fit a logistic regression with both L_1 penalty and L_2 penalty (Hui and Trevor, 2003). L_2 penalty is used to remedy the multi-collinearity problem, and it is a constraint on the sum of the square of the coefficients as expression 3.1. Here c is a positive constant.

$$L_2\ Penalty\ = \sum_{i=1}^{n} \beta_i^2\ \leq c \tag{3.1}$$

The L_1 and L_2 penalized maximum likelihood estimation solves the following optimization problem

$$min\left\{-\sum_{i=1}^{n}\left(y_i(\boldsymbol{\beta}^{\mathrm{T}}x_i) - \log\left(1 + e^{(\boldsymbol{\beta}^{\mathrm{T}}x_i)}\right)\right) + (1-\alpha)\sum_{i=1}^{p}|\beta_i| + \alpha\sum_{i=1}^{n}\beta_i^2\right\} \tag{3.2}$$

, where α is constant with value between 0 and 1. α is the tuning parameter for this minimization problem. The L_1 penalty might shrinkage some coefficients to exact zero while L_2 penalty does not. Using one dataset with non-converged fitted L_1 and L_2 regularized logistic regression model, when $\alpha = 0.4$, the fitted model reached convergence. So when there exists collinearity among the predictors, L_1 and L_2 regularized logistic regression is a good alternative to L_1 regularized logistic regression since it can remedy the collinearity problem with the L_2 penalty.

CHAPTER 4

Discussion and Future Work

4.1 Discussion

Lasso is a regression shrinkage and variable selection method that was proposed by Tibshirani in 1996. It adds a constraint on the model coefficients to achieve a sparse solution (Friedman et al., 2009). However, the Lasso does a poor job in the $p >> n$ case. A new regularization and variable selection method, elastic net, was introduced by Zou and Hastie in 2003. This method is viewed as a generalization of the lasso. It improves the Lasso and also is useful when number of predictors is much larger than the number of observations.

There are also many literatures to compare the Lasso with other variable selection methods for logistic regression. Fu (1998) compared the bridge regression, a special family of penalized regressions, with the Lasso, and concluded that the bridge regression performs well compared to the Lasso. However, Lasso is the well-developed shrinkage and variable selection method, in addition, a fast algorithm called coordinate descent for estimation of the logistic regression model was developed by Friedman, Hastie and Tibshirani. This allows a wider application of L_1 regularized logistic regression in practice.

4.2 Remaining Issues

First, our study only focuses on the selection of tuning parameter in two-class logistic regression. Further study can be conducted to evaluate the performance of BIC and CV for

choosing the tuning parameter in multinomial logistic regression models, or the case where the logit is a nonlinear function of predictors.

Second, in the case that number of predictors is much greater than the sample size, Lasso does not work well. Elastic net method is a better choice to select variables and estimate coefficients in this case. Last, although $glmnet$ package provides efficient procedures for fitting lasso regularization path for logistic regression, in the cases where the fitted models do not converge, a method to detect the non-convergence is desired.

Bibliography

Agresti, A. (2007). *An Introduction to Categorical Data Analysis.* New York: Wiley.

Akaike, H. (1973). Information Theory and an Extension of the Maximum Likelihood Principle. *Proc. 2nd Int. Symp. Info. Theory*, Ed. B. N. Petrov and F. Csaki, 267-281.

Burnham, K., and Anderson, D. (2002). *Model Selection and Multi-Model Inference*. New York: Springer.

Fan, J., and Li, R. (2006). Statistical Challenges with High Dimensionality: Feature Selection in Knowledge Discovery. *Proceedings of the International Congress of Mathematicians*, Vol. III, European Mathematical Society, Zurich, 595-622.

Fan, J., and Li, R. (2001). Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties. *Journal of the American Statistical Association*, 1348-1360.

Friedman, J., Hastie, T., and Tibshirani, R. (2009). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, 1-22.

Fu, W. (1998). Penalized Regressions: The Bridge versus the Lasso. *Journal of Computational and Graphical Statistics*, 397-416.

Hastie, T., Tibshirani, R., and Friedman, J. H. (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* New York: Springer.

Hosmer, D. W., and Lemeshow, S. (2000). *Applied Logistic Regression.* Chichester: Wiley.

Lee, S.-I. L., Abbeel, P., and Ng, A. (2006). Efficient L1 Regularized Logistic Regression. *American Association for Artificial Intelligence*.

Menard, S. W. (2002). Applied Logistic Regression. Sage Publications.

Ludden, T. M., Beal, S. L., and Sheiner, L. B. (1994). Comparison of the Akaike Information Criterion, the Schwarz Criterion and the F Test as Guides to Model Selection. *Journal of Pharmacokinetics and Biopharmaceutics*, 431-450.

Park, M. Y., and Hastie, T. (2007). L_1-regularization Path Algorithm for Generalized Linear Models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 659-677.

Peduzzi, P., Concato, J., Kemper, E., Holford, T.R. and Feinstein, A.R. (1996). A simulation study of the number of events per variable in logistic regression analysis. *Journal of clinical epidemiology*, 1373-1379.

Roth, V. (2004). The Generalized LASSO. *IEEE Transactions on Neural Networks*, 16-28.

Schwarz, Gideon E. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6 (2): 461–464.

Tang, W. H., and Tu, X. M. (2012). *Applied Categorical and Count Data Analysis.* Boca Raton: CRC.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 267-288.

Wang, H., Li, G. and Tsai, C. L. (2007a). Regression Coefficient and autoregressive orrder shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69, 63-78.

Yang, Y. (2005). Can the Strengths of AIC and BIC Be Shared? A Conflict between Model Indentification and Regression Estimation. *Biometrika*, 937-950.

Zellner, D., Keller, F., and Zellner, G. E. (2004). Variable Selection in Logistic Regression Models. *Communications in Statistics - Simulation and Computation*, 787-805.

Zou, H., and Hastie, T. (2005). Regularization and Variable Selection via the Elastic Net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 301-320.

Appendix

R Code

In this appendix, we give the R code for simulating the 200 datasets based on the model setup as $p = 10$, vector of coefficients as $\boldsymbol{\beta_2} = (1, 1, 1, 1, 0.5, 0.5, 0.5, 0.5, 0, 0)$, $n = 100$ and $\rho = 0.2$ and computing the average bias, TPR and FPR of the 200 fitted models.

```
### Package glmnet was used to fit the L1 regularized logistic regression model
library(glmnet)

### Simulated 200 datasets for each model setup
rpt=200
### Number of predictors
p=10
### Number of important predictors
p1=0.2*p
### Number of unimportant predictors
p0=p-p1
### Sample size
n=100
### mean of the predictors
mu=rep(0,p)
### Correlation structure of the predictors
sigma=diag(1,p)
for (i in 1:p){
 for (j in 1:p){
  sigma[i,j] <- 0.8^(abs(i-j))
 }
}
### vector of coefficients of predictors
beta=c(rep(1,p1/2),rep(0.5,p1/2),rep(0,p0))

### Storage for the bias when using BIC
Bias_BIC=rep(NA,rep=rpt)
### Storage for the bias when using CV
```

```
Bias_CV =rep(NA,rep=rpt)
### Storage for the TPR when using BIC
TPR_BIC =rep(NA,rep=rpt)
### Storage for the FPR when using BIC
FPR_BIC =rep(NA,rep=rpt)
### Storage for the TPR when using CV
TPR_CV  =rep(NA,rep=rpt)
### Storage for the FPR when using CV
FPR_CV  =rep(NA,rep=rpt)

for (i in 1:rpt){
  ### Simulate matrix of predictors x
  x <- mvrnorm(n,mu,sigma)
  ### Simulate values of response y
  prob <- 1/(1+exp(-(x%*%beta)))
  y <- rbinom(n,1,prob)

  ### fit L1 regularized logistic regression models
  fit.BIC <- glmnet(x,y,family="binomial")
  ### calculate the BIC for each fitted model
  BIC <- deviance(fit.BIC)+(fit.BIC$df+1)*(log(n))
  ### Get the coefficients estimation of the fitted model with min BIC
  coef.BIC <- as.matrix(coef(fit.BIC)[2:(p+1),which.min(BIC)])
  ### Get the coefficients estimates of the true important predictors
  b1.BIC <- coef.BIC[1:p1]
  ### Get the coefficients estimates of the predictors with true coefficients 1
  b11.BIC <- coef.BIC[1:(p1/2)]
  ### Get the coefficients estimates of the predictors with true coefficients 0.5
  b12.BIC <- coef.BIC[((p1/2)+1):p1]
  ### Get the coefficients estimates of the predictors with true coefficients 0
  b0.BIC <- coef.BIC[(p1+1):p]
  ### Calculate the Bias of the fitted model with min BIC
  Bias_BIC[i] <- sum(abs(1-b11.BIC))+sum(abs(0.5-b12.BIC))+sum(abs(0-b0.BIC))
  ### Calculate the TPR of the fitted model with min BIC
  TPR_BIC[i] <- length((abs(b1.BIC)>0.01)[(abs(b1.BIC)>0.01)==TRUE])/p1
  ### Calculate the TPR of the fitted model with min BIC
  FPR_BIC[i] <- length((abs(b0.BIC)>0.01)[(abs(b0.BIC)>0.01)==TRUE])/p0

  ### fit L1 regularized logistic regression models and
  ### return the tuning parameter with min cross validated error
  fit.CV <- cv.glmnet(x,y,family="binomial")
```

```
  ### Get the coefficients estimation of the selected model
  coef.CV <- as.matrix(coef(fit.CV)[2:(p+1),])
  ### Get the coefficients estimates of the predictors with true important predictors
  b1.CV <- coef.CV[1:p1]
  ### Get the coefficients estimates of the predictors with true coefficients 1
  b11.CV <- coef.CV[1:(p1/2)]
  ### Get the coefficients estimates of the predictors with true coefficients 0.5
  b12.CV <- coef.CV[((p1/2)+1):p1]
  ### Get the coefficients estimates of the predictors with true coefficients 0
  b0.CV <- coef.CV[(p1+1):p]
  ### Calculate the Bias of the selected model
  Bias_CV[i] <- sum(abs(1-b11.CV))+sum(abs(0.5-b12.CV))+sum(abs(0-b0.CV))
  ### Calculate the TPR of the selected model
  TPR_CV[i] <- length((abs(b1.CV)>0.01)[(abs(b1.CV)>0.01)==TRUE])/p1
  ### Calculate the FPR of the selected model
  FPR_CV[i] <- length((abs(b0.CV)>0.01)[(abs(b0.CV)>0.01)==TRUE])/p0
}

### Result of BIC
### Mean of the 200 Biases
Bias.BIC <- mean(Bias_BIC)
### Standard deviation of the 200 Biases
Std.Bias.BIC <- sd(Bias_BIC)
### Mean of the 200 Adjusted Biases
Adj.Bias.BIC <- mean(Bias_BIC/p1)
### Standard deviation of the 200 Adjusted Biases
Std.Adj.Bias.BIC <- sd(Bias_BIC/p1)
### Mean of the 200 TPRs
TPR.BIC <- mean(TPR_BIC)
### Standard deviation of the 200 TPRs
Std.TPR.BIC <- sd(TPR_BIC)
### Mean of the 200 FPRs
FPR.BIC <- mean(FPR_BIC)
### Standard deviation of the 200 FPRs
Std.FPR.BIC <- sd(FPR_BIC)

### Result of CV
### Mean of the 200 Biases
Bias.CV <- mean(Bias_CV)
### Standard deviation of the 200 Biases
Std.Bias.CV <- sd(Bias_CV)
### Mean of the 200 Adjusted Biases
Adj.Bias.CV <- mean(Bias_CV/p1)
### Standard deviation of the 200 Adjusted Biases
Std.Adj.Bias.CV <- sd(Bias_CV/p1)
### Mean of the 200 TPRs
```

```
TPR.CV <- mean(TPR_CV)
### Standard deviation of the 200 TPRs
Std.TPR.CV <- sd(TPR_CV)
### Mean of the 200 FPRs
FPR.CV <- mean(FPR_CV)
### Standard deviation of the 200 FPRs
Std.FPR.CV <- sd(FPR_CV)

### END ###
```