รายงานการวิจัยฉบับสมบูรณ์

ตัวแบบสำหรับพยากรณ์มะเร็งเต้านมด้วยวิธีการวิเคราะห์การถดถอยโลจิสติกส์ริดจ์

Modelling for Prediction Breast Cancer by Logistic Ridge Regression

นางสาวอัชฌา อระวีพร

ได้รับทุนสนับสนุนงานวิจัยจากเงินรายได้ ประจำปีงบประมาณ 2561
คณะวิทยาศาสตร์
สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

รายงานการวิจัยฉบับสมบูรณ์

ตัวแบบสำหรับพยากรณ์มะเร็งเต้านมด้วยวิธีการวิเคราะห์การถดถอยโลจิสติกส์ริดจ์

Modelling for Prediction Breast Cancer by Logistic Ridge Regression

นางสาวอัชฌา อระวีพร

ได้รับทุนสนับสนุนงานวิจัยจากเงินรายได้ ประจำปีงบประมาณ 2561
คณะวิทยาศาสตร์
สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

Research Title : Modelling for Prediction Breast Cancer by Logistic Ridge Regression

Researcher : Autcha Araveeporn

Department : Statistics

Faculty : Science

ABSTRACT

The goal of this research is to estimate the parameter of logistic regression model. The coefficient parameter is evaluated by maximum likelihood, ridge regression, markov chain monte carlo methods. The logistic regression is considered the correlation between binary dependent variable and 2, 3, and 4 independent variables which is generated from normal distribution, contaminated normal distribution, and t distribution. The maximum likelihood estimator is estimated by differential the log likelihood function with respect to the coefficients. Ridge regression is to choose the unknown ridge parameter by cross-validation, so ridge estimator is evaluated on a form of maximum likelihood method by adding ridge parameter. The markov chain monte carlo estimator can approximate from Gibbs sampling algorithm by the posterior distribution based on a probability distribution and prior probability distribution. The performance of these method is compare by percentage of predicted accuracy value. The results are found that ridge regression are satisfied when the independent variables are simulated from normal distribution, and the maximum likelihood outperforms on the other distributions.

The logistic regression model by penalized regression analysis consisted of ridge regression, lasso, and elastic net method. The logistic regression is considered between binary dependent variable and 3 and 5 independent variables. The independent variables are generated form normal distribution, contaminated normal distribution, and t distribution on correlation coefficient at 0.1, 0.5, and 0.99 or called multicollinearity problem. The maximum likelihood estimator is the classical method by differential the log likelihood function with respect to the coefficients. Ridge regression is to choose the unknown ridge parameter by cross-validation, so ridge estimator is evaluated by adding ridge parameter on penalty term. Lasso (least absolute shrinkage and selection operator)

is added the penalty term on scales sum of the absolute value of the coefficients. The elastic net can be mixed between ridge regression and lasso on the penalty term. The criterion of these method is compare by percentage of predicted accuracy value. The results are found that lasso are satisfied when the independent variables are simulated from normal and t distribution in most cases, and the lasso outperforms on the contaminated normal distribution.

Keywords: Elastic Net, Lasso, Markov Chain Monte Carlo, Maximum Likelihood, Ridge Regression

CONTENTS

Chapter 1

Logistic Regression model without correlation

1.1 Introduction

Multiple regression analysis is to learn about the association between several independent variables and dependent variable for create the multiple regression function which is used to predict and estimate dependent given the independent variables. The assumption of multiple regression analysis involves checking to make sure that the data can actually using multiple regression such as the dependent and independent variables on a continuous scale. But the dependent variable is a categorical variable based on discrete scale, the logistic regression can be used to construct the model for forecasting dependent variable as the multiple regression analysis.

Logistic regression analysis studies the relationship between a categorical dependent variable and a set of independent variables by estimating probabilities using a logistic function. When the dependent variable has only two values, for example "dead" vs. "alive" or "win" vs. "loss", it's called binary logistics regression. For multinomial logistic regression, the dependent variable has more two values. The objective of logistic regression is to report the model for predicted values from independent variables when the independent variable shows the continuous variable and no multicollinearity problem.

The maximum likelihood method is a well known method for estimating parameter of statistical model given observation. The estimator is estimated by maximize the likelihood function given the parameter. Lee, Silvapulle [1] observed that a ridge type estimator is at least as good as the maximum likelihood estimator in terms of total and prediction mean squared error criteria. Duffy, Santner [2] considered the maximization of the log-likelihood function with a penalty value or called ridge parameter. The use of ridge regressin is developed from regression analysis by choosing the ridge parameter. Cessie, Houwelingen [3] proposed ridge parameter in logistic regression by using cross-validation. The Bayes' method uses both a probability distribution and prior probability distribution to approach a posterior probability distribution (Bradley, Thomas, [5]). Then it is difficult to demonstrate a posterior distribution from a probability distribution and prior probability distribution. However, the

Markov Chain Monte Carlo (MCMC) method (Gilks, Richardson, Speigelhalter [6]) can approximate the estimator from Gibbs sampling algorithm (Geman, Geman [7]) based on the posterior distribution.

In this paper, we focus to estimate the coefficient parameter on logistic regression when the dependent variable occurs in binary data. The maximum likelihood, ridge regression, MCMC methods are used to improve the parameter estimates by further predictions. Various methods to determine the parameter estimation are discussed in method for estimation. In simulation study, logistic regression is presented the detail of simulation data based on independent variable, and the results are by percentage of predicted accuracy.

1.2 Logistic Regression Model

The logistic regression consisted of binary dependent variable (Y_i) and independent variables ($\underset{\sim}{x}_i$), where $\underset{\sim}{x}_i = x_{1i}, x_{2i}, ..., x_{ki}$, k is a number of independent variable, and $i = 1, 2, ..., n$ is a number of observed data. The most idea is to let $p(\underset{\sim}{x}_i)$ be a probability function in term of linear function of $\underset{\sim}{x}_i$. Let $\log p(\underset{\sim}{x}_i)$ be a linear function of $\underset{\sim}{x}_i$, so that changing an independent variables multiplies the probability. The easiest modification of $\log p(\underset{\sim}{x}_i)$ which has an unbound range is the logistic transformation as $\log \frac{p(\underset{\sim}{x}_i)}{1-p(\underset{\sim}{x}_i)}$.

Formally, the logistic regression model is shown that

$$\log \frac{p(\underset{\sim}{x}_i)}{1-p(\underset{\sim}{x}_i)} = \beta_0 + x_{1i}\beta_1 + x_{2i}\beta_2 + ... + x_{ki}\beta_k, \tag{1.1}$$

where k is a number of independent variable, and $i = 1, 2, ..., n$ is a number of observed data.

The probability function follows the logistic regression model

$$p(\underset{\sim}{x}_i) = \frac{e^{\beta_0 + x_{1i}\beta_1 + x_{2i}\beta_2 + ... + x_{ki}\beta_k}}{1 + e^{\beta_0 + x_{1i}\beta_1 + x_{2i}\beta_2 + ... + x_{ki}\beta_k}} = \frac{1}{1 + e^{-(\beta_0 + x_{1i}\beta_1 + x_{2i}\beta_2 + ... + x_{ki}\beta_k)}}. \tag{1.2}$$

The classification rate is predicted by $Y_i = 1$, when $p(\underset{\sim}{x}_i) \geq 0.5$, and $Y_i = 0$, when $p(\underset{\sim}{x}_i) < 0.5$.

1.3 Method for Estimation

1.3.1 Maximum likelihood Method

Logistic regression predicts the probability function by classification on dependent variable in 2 classes. The likelihood function is then

$$L(\beta_0, \beta_1, ..., \beta_k) = \prod_{i=1}^{n} p(\underset{\sim}{x}_i)^{Y_i} (1 - p(\underset{\sim}{x}_i))^{1-Y_i}. \quad (1.3)$$

The log likelihood function turns into sum as :

$$\begin{aligned}
\log L(\beta_0, \beta_1, ..., \beta_k) &= \sum_{i=1}^{n} \left[Y_i \log p(\underset{\sim}{x}_i) + (1 - Y_i) \log \{1 - p(\underset{\sim}{x}_i)\} \right] \\
&= \sum_{i=1}^{n} \left[Y_i \log p(\underset{\sim}{x}_i) + \log \{1 - p(\underset{\sim}{x}_i)\} - Y_i \log \{1 - p(\underset{\sim}{x}_i)\} \right] \\
&= \sum_{i=1}^{n} \left[\log \{1 - p(\underset{\sim}{x}_i)\} \right] + \sum_{i=1}^{n} Y_i \log \frac{p(\underset{\sim}{x}_i)}{1 - p(\underset{\sim}{x}_i)} \\
&= \sum_{i=1}^{n} \left[-\log 1 + e^{\beta_0 + x_{1i}\beta_1 + x_{2i}\beta_2 + ... + x_{ki}\beta_k} \right] \\
&\quad + \sum_{i=1}^{n} Y_i \left(\beta_0 + x_{1i}\beta_1 + x_{2i}\beta_2 + ... + x_{ki}\beta_k \right).
\end{aligned}$$

The maximum likelihood estimator is approximated by differential the log likelihood function with respect to the parameters, and set the derivatives equal to zero. The log likelihood takes the derivatives with respect to one parameter of $\beta_j, j = 1, 2, ..., k$ by

$$\begin{aligned}
\frac{\partial \log L(\beta_0, \beta_1, ..., \beta_k)}{\partial \beta_j} &= -\sum_{i=1}^{n} \frac{1}{1 + e^{\beta_0 + x_{1i}\beta_1 + x_{2i}\beta_2 + ... + x_{ki}\beta_k}} e^{\beta_0 + x_{1i}\beta_1 + x_{2i}\beta_2 + ... + x_{ki}\beta_k} \underset{\sim}{x}_i + \sum_{i=1}^{n} Y_i \underset{\sim}{x}_i \\
&= \sum_{i=1}^{n} \left(Y_i - p(\underset{\sim}{x}_i) \right) \underset{\sim}{x}_i.
\end{aligned}$$

Above equation can not to set as zero and solve exactly, so we can approximate the parameter by numerical method as $\underset{\sim}{\hat{\beta}}_{ML} = (\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_k)$.

1.3.2 Ridge Regression Method

Hoerl, Kennard [8] proposed the method to solve the problems when the independent variables have multicorllinearity and get the minimum mean square error or called ridge regression method. From the multiple linear regression analysis,

the least squares estimation is normally used to approximate the parameter of multiple linear regression model by

$$\underset{\sim}{\hat{\beta}} = (X'X)^{-1}X'\underset{\sim}{Y}, \tag{1.4}$$

where $\underset{\sim}{\hat{\beta}}$ is a vector of unknown parameter as $(\beta_0, \beta_1, ..., \beta_k)$, X is an independent variable matrix by the $(k+1)$ columns and n rows, and $\underset{\sim}{Y}$ is a vector of dependent variables as $(Y_1, Y_2, ..., Y_n)$.

The idea of ridge regression estimation is a procedure based on adding ridge parameter as small positive quantities to the diagonal of matrix $X'X$. The ridges regression estimators are estimated by

$$\underset{\sim}{\hat{\beta}}_R = (X'X + \lambda I)^{-1}X'\underset{\sim}{Y}, \lambda > 0. \tag{1.5}$$

It can be used to obtain an estimated parameter with smaller mean square error. The cross-validation method is chosen to find the smallest ridge parameter (λ).

1.3.3 Markov Chain Monte Carlo Method

The Markov Chain Monte Carlo (MCMC) (Gamerman [9]) method is operated by sequentially sampling parameter values from a Markov Chain at stationary distribution which is desired from posterior distribution. The Gibbs sampling (Gelfand, Hills, Racine-Poon, Smith [10]) is an algorithm for MCMC computing. We carry out the WinBUGS Program (Lunn, Spiegelhalter, Thomas, Best [11]) to obtain the estimating estimator from the posterior distribution function based on MCMC process.

The logistic regression is used the logit model following

$$\text{logit}(p(\underset{\sim}{x}_i)) = \log\frac{p(\underset{\sim}{x}_i)}{1-p(\underset{\sim}{x}_i)} = \beta_0 + x_{1i}\beta_1 + x_{2i}\beta_2 + ... + x_{ki}\beta_k, \tag{1.6}$$

and we can model the probability of each subject i as a Bernoulli distribution and the prior distribution is considered the normal distribution. The likelihood function can be implied as

$$\text{L}(\beta_0, \beta_1, ..., \beta_k) \propto \prod_{i=1}^{n} p(\underset{\sim}{x}_i)^{Y_i}(1-p(\underset{\sim}{x}_i))^{1-Y_i},$$

and the posterior distribution function is the product of the joint prior distribution, the likelihood function in terms of

$$p(\beta_0,\beta_1,...,\beta_k \mid Y_i,\underset{\sim}{x}_i) \propto p(\beta_0,\beta_1,...,\beta_k)\prod_{i=1}^{n} p(\underset{\sim}{x}_i)^{Y_i}(1-p(\underset{\sim}{x}_i))^{1-Y_i}. \quad (1.7)$$

The Gibbs sampling algorithm of logistic regression model is specified in hierarchical model following

$$\begin{aligned} y_i &\sim Bernoulli(p_i) \\ \text{Logit}(p_i) &= \beta_0 + x_{1i}\beta_1 + \ldots + x_{ki}\beta_k \\ \beta_0 &\sim Normal(0, 0.0001) \\ \beta_1 &\sim Normal(0, 0.0001) \\ &\ldots \\ \beta_k &\sim Normal(0, 0.0001). \end{aligned}$$

The MCMC samples of $\hat{\underset{\sim}{\beta}}_{MCMC} = (\hat{\beta}_0,\hat{\beta}_1,...,\hat{\beta}_k)$ obtain the posterior mean as coefficient estimators of logistic regression by $\hat{\underset{\sim}{\beta}}_{MCMC}$.

1.4 Simulation Study

In this section, we show the detail of a simulated data that we conducted in order to compare the performance of maximum likelihood, ridge regression, and markov chain monte carlo methods for logistic regression. To simulate data, we generated data independent variables in class of 2,3, and 4 variables based on normal distribution at mean zero and variance one, contaminated normal distribution at contaminated data with 5 and 10 percent (p = 0.05, 0.1) on variance of nine, and T distribution at 3 degree of freedom by R statistical software. The sample size is set as 30, 50, and 100 with 500 times in each cases. The set of coefficient parameter on logistic regression $(\beta_0,\beta_1,\beta_2)$, $(\beta_0,\beta_1,\beta_2,\beta_3)$, and $(\beta_0,\beta_1,\beta_2,\beta_3,\beta_4)$ is defined as constant value based on respective independent variable.

The example of two independent variables, the probability function follows the logistic regression model

$$p(\underset{\sim}{x}_i) = \frac{1}{1+e^{-(\beta_0+x_{1i}\beta_1+x_{2i}\beta_2)}}.$$

If $p(\underset{\sim}{x}_i) \geq 0.5$, the dependent variables will be define $Y_i = 1$, and $Y_i = 0$, when $p(\underset{\sim}{x}_i) < 0.5$.

After the estimating parameter of 3 methods, we obtain the coefficient parameter as $(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2)$, then we approximate the probability function by

$$\hat{p}(\underset{\sim}{x}_i) = \frac{1}{1+e^{-(\hat{\beta}_0+x_{1i}\hat{\beta}_1+x_{2i}\hat{\beta}_2)}}.$$

The dependent values are predicted by $\hat{Y}_i = 1$ when $\hat{p}(\underset{\sim}{x}_i) \geq 0.5$, and $\hat{Y}_i = 0$ when $\hat{p}(\underset{\sim}{x}_i) < 0.5$.

The confusion matrix is a table that is often used to describe the performance of a classification model on a set of predicted data for which the actual data are known following on Table 1.1. The predicted accuracy is computed by

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}.$$

Table 1.1 : The confusion matrix of actual data (Y_i) and predicted data ($\hat{Y}_i$)

Predicted Data	Actual Data	
	$Y_i = 1$	$Y_i = 0$
$\hat{Y}_i = 1$	True Positive (TP)	False Positive (FP)
$\hat{Y}_i = 0$	False Negative (FN)	True Negative (TN)

1.5 Results

The estimating coefficient of logistic regression model is obtained from the maximum likelihood (ML), ridge regression (Ridge), and markov chain monte carlo (MCMC) methods which transformed to logit model and classified to binary dependent variable. Table 1.2-1.4 present the average percentage of predicted accuracy on previous methods. The maximizing percentage are illustrated the performance of these methods.

Table 1.2 : The average percentage of predicted accuracy of maximum likelihood (ML), ridge regression (Ridge), and markov chain monte carlo (MCMC) methods on 2 independent variables

Distributions	Sample size	ML	Ridge	MCMC
Normal	n=30	93.56	**95.70**	49.90
	n=50	95.82	**96.04**	49.44
	n=100	**97.72**	96.28	49.62
Contaminated Normal (p=0.05)	n=30	**93.54**	80.52	51.10
	n=50	**95.08**	78.00	48.98
	n=100	**97.72**	76.88	50.94
Contaminated Normal (p=0.1)	n=30	**93.52**	77.27	49.44
	n=50	**95.65**	75.85	48.92
	n=100	**97.67**	77.17	50.37
t	n=30	**93.66**	89.34	51.76
	n=50	**95.83**	89.30	51.26
	n=100	**97.57**	90.21	50.34

Table 1.3 : The average percentage of predicted accuracy maximum likelihood (ML), ridge regression (Ridge), and markov chain monte carlo (MCMC) methods on 3 independent variables

Distributions	Sample size	ML	Ridge	MCMC
Normal	n=30	91.90	**95.95**	49.57
	n=50	94.72	**96.90**	49.65
	n=100	97.20	**97.24**	50.38
Contaminated Normal (p=0.05)	n=30	**91.63**	80.78	51.08
	n=50	**94.52**	78.77	48.89
	n=100	**97.06**	79.70	49.50
Contaminated Normal (p=0.1)	n=30	**94.54**	78.12	49.36
	n=50	**94.54**	79.30	48.53
	n=100	**97.02**	81.59	50.11
t	n=30	**91.70**	91.20	50.38
	n=50	**94.77**	91.20	49.84
	n=100	**97.00**	92.58	49.66

Table 1.4 : The average percentage of predicted accuracy of maximum likelihood (ML), ridge regression (Ridge), and markov chain monte carlo (MCMC) methods on 4 independent variables

Distributions	Sample size	ML	Ridge	MCMC
Normal	n=30	90.00	**95.42**	48.46
	n=50	**96.73**	96.58	50.58
	n=100	96.57	**97.30**	50.34
Contaminated Normal (p=0.05)	n=30	**89.90**	81.63	48.74
	n=50	**93.42**	80.66	49.39
	n=100	**96.31**	81.36	49.74
Contaminated Normal (p=0.1)	n=30	**85.99**	80.78	50.70
	n=50	**93.34**	82.19	50.60
	n=100	**96.30**	84.13	50.19
t	n=30	90.19	**90.89**	48.99
	n=50	**93.59**	92.06	51.46
	n=100	**96.46**	92.90	49.27

From Table 1.2-1.4, the percentage of ridge regression method is a maximum average percentage values for all sample size when the independent variable is simulated from normal distribution. For maximum likelihood method, contaminated normal and t distribution via independent variables appear the maximum average percentage values. The remaining Table 4 of four independent variables, the results are similar the Table 1.2-1.4 except n=50 with normal distribution and n=30 with t distribution.

1.6 Conclusion

In this research, we generated independent variables from normal distribution, contaminated normal distribution, and t distribution. The maximum likelihood, ridge regression, and markov chain monte carlo methods are used to estimate parameter on ridge regression and classified the binary dependent variable. The ridge regression is a good performance when the independent variable is presented on the normal distribution in most cases. Therefore, the maximum likelihood method is a good fit when the independent variables is played on contaminated normal distribution and t distribution.

Chapter 2

Logistic Regression with Correlation

2.1 Introduction

Multiple regression analysis is to learn about the association between several independent variables and dependent variable for create the multiple regression function which is develop to carry out parameter estimation of dependent variable based on the independent variables. The assumption of multiple regression analysis involves for checking to make sure that the data can actually using multiple regression such as the dependent and independent variables on a continuous scale. But the dependent variable is a categorical variable based on discrete scale, the logistic regression can be used to construct the model for forecasting dependent variable as the multiple regression analysis.

Logistic regression analysis studies the relationship between a categorical dependent variable and a set of independent variables by estimating probabilities using a logistic function. When the dependent variable has only two values, for example "dead" vs. "alive" or "win" vs. "loss", it's called binary logistics regression. For multinomial logistic regression, the dependent variable has more two values. The objective of logistic regression is to report the model for estimated values from independent variables when the independent variable shows the continuous variable with multicollinearity problem.

The maximum likelihood method is a well known method for estimating parameter of statistical model given observation. The estimator is estimated by maximize the likelihood function given the parameter. Lee, Silvapulle [1] observed that a ridge type estimator is at least as good as the maximum likelihood estimator in terms of total and prediction mean squared error criteria. Duffy, Santner [2] considered the maximization of the log-likelihood function with a penalty value or called ridge parameter. The use of ridge regression is developed from regression analysis by choosing the ridge parameter. Cessie, Houwelingen [3] proposed ridge parameter in logistic regression by using cross-validation. Tibshirani [4] produced interpretable models like subset selection and exhibited the stability of ridge regression called lasso. The concept of lasso is minimized the residual sum of squares to the sum of absolute value of the coefficients being less than constant. Zou and Hastie [12] proposed a new regularization and variable selection method called elastic net. The elastic

net idea is mixed between ridge regression and lasso to select groups of correlated variables.

In this paper, we focus to estimate the coefficient parameter on logistic regression when the dependent variable occurs in binary data. The maximum likelihood, ridge regression, lasso, and elastic net methods are used to improve the parameter estimates by further estimations. Various methods to determine the parameter estimation are discussed in method for estimation. In simulation study, logistic regression is presented the detail of simulation data based on independent variable, and the results are by percentage of predicted accuracy. For actual data, we used the breast cancer databases to estimate parameter by penalized regression analysis on logistic regression model.

2.2 Materials and methods

The logistic regression consisted of binary dependent variable (Y_i) and independent variables ($\underset{\sim}{x}_i$), where $\underset{\sim}{x}_i = x_{i1}, x_{i2}, ..., x_{ik}$, k is a number of independent variable, and $i = 1, 2, ..., n$ is a number of observed data. The most idea is to let $p(\underset{\sim}{x}_i)$ be a probability function in term of linear function of $\underset{\sim}{x}_i$. Let $\log p(\underset{\sim}{x}_i)$ be a linear function of $\underset{\sim}{x}_i$, so that changing an independent variables multiplies the probability. The easiest modification of $\log p(\underset{\sim}{x}_i)$ which has an unbound range is the logistic transformation as $\log \frac{p(\underset{\sim}{x}_i)}{1-p(\underset{\sim}{x}_i)}$. Formally, the logistic regression model is shown that

$$\log \frac{p(\underset{\sim}{x}_i)}{1-p(\underset{\sim}{x}_i)} = \beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2 + ... + x_{ik}\beta_k, \qquad (2.1)$$

where k is a number of independent variable, and $i = 1, 2, ..., n$ is a number of observed data.

From (2.1), the probability function follows the logistic regression model

$$p(\underset{\sim}{x}_i) = \frac{e^{\beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2 + ... + x_{ik}\beta_k}}{1 + e^{\beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2 + ... + x_{ik}\beta_k}} = \frac{1}{1 + e^{-(\beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2 + ... + x_{ik}\beta_k)}}. \qquad (2.2)$$

The classification rate is predicted by $Y_i = 1$, when $p(\underset{\sim}{x}_i) \geq 0.5$, and $Y_i = 0$, when $p(\underset{\sim}{x}_i) < 0.5$.

2.2.1 Maximum likelihood Method

Logistic regression predicts the probability function by classification on dependent variable in 2 classes. The likelihood function is then

$$\mathrm{L}(\beta_0,\beta_1,...,\beta_k) = \prod_{i=1}^{n} p(\underset{\sim}{x}_i)^{Y_i}(1-p(\underset{\sim}{x}_i))^{1-Y_i}. \tag{2.3}$$

The log likelihood function turns into sum as :

$$\begin{aligned}\log \mathrm{L}(\beta_0,\beta_1,...,\beta_k) &= \sum_{i=1}^{n}\left[Y_i \log p(\underset{\sim}{x}_i)+(1-Y_i)\log\{1-p(\underset{\sim}{x}_i)\}\right]\\ &= \sum_{i=1}^{n}\left[Y_i \log p(\underset{\sim}{x}_i)+\log\{1-p(\underset{\sim}{x}_i)\}-Y_i\log\{1-p(\underset{\sim}{x}_i)\}\right]\\ &= \sum_{i=1}^{n}\left[\log\{1-p(\underset{\sim}{x}_i)\}\right]+\sum_{i=1}^{n}Y_i\log\frac{p(\underset{\sim}{x}_i)}{1-p(\underset{\sim}{x}_i)}\\ &= \sum_{i=1}^{n}\left[-\log 1+e^{\beta_0+x_{i1}\beta_1+x_{i2}\beta_2+...+x_{ik}\beta_k}\right]\\ &\quad +\sum_{i=1}^{n}Y_i\left(\beta_0+x_{i1}\beta_1+x_{i2}\beta_2+...+x_{ik}\beta_k\right).\end{aligned}$$

The maximum likelihood estimator is approximated by differential the log likelihood function with respect to the parameters, and set the derivatives equal to zero. The log likelihood takes the derivatives with respect to one parameter of $\beta_j, j=1,2,...,k$ by

$$\begin{aligned}\frac{\partial \log \mathrm{L}(\beta_0,\beta_1,...,\beta_k)}{\partial \beta_j} &= -\sum_{i=1}^{n}\frac{1}{1+e^{\beta_0+x_{i1}\beta_1+x_{i2}\beta_2+...+x_{ik}\beta_k}}e^{\beta_0+x_{i1}\beta_1+x_{i2}\beta_2+...+x_{ik}\beta_k}\underset{\sim}{x}_i+\sum_{i=1}^{n}Y_i\underset{\sim}{x}_i\\ &= \sum_{i=1}^{n}\left(Y_i-p(\underset{\sim}{x}_i)\right)\underset{\sim}{x}_i.\end{aligned}$$

Above equation can not to set as zero and solve exactly, so we can approximate the parameter by numerical method as $\underset{\sim}{\hat{\beta}}_{ML} = (\hat{\beta}_0,\hat{\beta}_1,...,\hat{\beta}_k)$.

2.2.2 Ridge Regression Method

Penalized regression methods keep all the predictor variables in the model but constrain the regression coefficients by shrinking them toward zero. If the amount of shrinkage is large enough, these methods can also perform variable selection by shrinking some coefficients to zero. This objective function are formulated in the constrained form following

$$\hat{\underset{\sim}{\beta}} = \arg\min \sum_{i=1}^{n} \left(Y_i - (\sum_{j=1}^{k} x_{ij}\beta_j) \right)^2 + P_\lambda(\beta).$$

The solution for the vector logistic regression coefficients, $\hat{\underset{\sim}{\beta}}$, is minimizing the objective function to a penalty function on the logistic regression coefficients, $P_\lambda(\beta)$.

Hoerl, Kennard [8] proposed the method to solve the problems when the independent variables have multicorllinearity and get the minimum mean square error or called ridge regression method. This objective function of ridge regression is

$$\hat{\underset{\sim}{\beta}}_R = \arg\min \sum_{i=1}^{n} \left(Y_i - (\sum_{j=1}^{k} x_{ij}\beta_j) \right)^2 + \lambda \sum_{j=1}^{k} \beta_j^2. \quad (2.4)$$

Where the penalty function is $\lambda \sum_{j=1}^{k} \beta_j^2$, the tuning parameter serves λ to control the relative impact of these two terms on the logistic regression coefficient estimates. It can be used to obtain an estimated parameter with smaller mean square error. The cross-validation method is chosen to find the smallest ridge parameter (λ).

2.2.3 Lasso Method

Tibshirani [4] proposed a new technique called lasso, from 'least absolute shrinkage and selection operator'. The lasso does both continuous shrinkage and automatic variable selection simultaneously. It shrinks some coefficients and sets other to 0, and tries to retain the good features of both subset selection and ridge regression. The objective function of lasso is

$$\hat{\underset{\sim}{\beta}}_L = \arg\min \sum_{i=1}^{n} \left(Y_i - (\sum_{j=1}^{k} x_{ij}\beta_j) \right)^2 + \lambda \sum_{j=1}^{k} |\beta_j|. \quad (2.5)$$

Where the penalty function is $\lambda \sum_{j=1}^{k} |\beta_j|$, the tuning parameter λ determines the $\hat{\underset{\sim}{\beta}}$ that shrunk towards 0. If λ is equal to 0, then there is no shrinkage at all, whereas if λ is huge, the function can be minimized by shrinking all the parameters back towards 0. The cross-validation method is used to try out different values of λ, while the other data are used to assess the predictive performance of models of the different complexities. Lasso estimators for all values of λ can be computed through a modification of the LARS algorithm [7].

2.2.4 Elastic Net Method

Zou and Hastie [13] proposed a new regularization technique that called the elastic net. Similar to the lasso, the elastic net simultaneously does automatic variable selection and continuous shrinkage, and it can select groups of correlated variables. The objective function of elastic net is

$$\hat{\beta}_{E} = \arg\min \sum_{i=1}^{n}\left(Y_i - (\sum_{j=1}^{k} x_{ij}\beta_j)\right)^2 + \lambda_1 \sum_{j=1}^{k} \beta_j^2 + \lambda_2 \sum_{j=1}^{k} |\beta_j|. \quad (2.6)$$

Elastic net penalty is a convex combination of the ridge and lasso penalty. When $\lambda_1 = 0$, the elastic net becomes simple ridge regression. The tuning parameter λ_1 and λ_2 are selected by cross-validation [14].

2.3 Simulation Study

A simulated data conducted in order to compare the performance of maximum likelihood, ridge regression, lasso, elastic net for logistic regression. To simulate data, we generated data independent variables in class of 3 and 5 variables based on normal distribution at mean zero and variance one, contaminated normal distribution at contaminated data with 5 and 10 percent (p = 0.05, 0.1) on variance 9, and t distribution at 3 degree of freedom by R statistical software. The correlation coefficient of two variable is 0.1, 0.5, and 0.99 which is generated on the multivariate normal distribution. The sample size is set as 30, 50, and 100 with 500 times in each case. The set of coefficient parameter on logistic regression $(\beta_0, \beta_1, \beta_2, \beta_3)$ and $(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5)$ is defined as constant value based on respective independent variable. For three and five independent variables, the probability function follows the logistic regression model as

$$p(\underset{\sim}{x}_i) = \frac{1}{1+e^{-(\beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2 + x_{i3}\beta_3)}} \text{ and } p(\underset{\sim}{x}_i) = \frac{1}{1+e^{-(\beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2 + x_{i3}\beta_3 + x_{i4}\beta_4 + x_{i5}\beta_5)}}.$$

If $p(\underset{\sim}{x}_i) \geq 0.5$, the dependent variables will be define $Y_i = 1$, and $Y_i = 0$, when $p(\underset{\sim}{x}_i) < 0.5$.

After the estimating parameter of 4 methods, we obtain the coefficient parameter as $(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2)$ and $(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_4, \hat{\beta}_5)$ then we approximate the probability function by

$$\hat{p}(\underset{\sim}{x}_i) = \frac{1}{1+e^{-(\hat{\beta}_0 + x_{i1}\hat{\beta}_1 + x_{i2}\hat{\beta}_2 + x_{i3}\hat{\beta}_3)}} \text{ and } \hat{p}(\underset{\sim}{x}_i) = \frac{1}{1+e^{-(\hat{\beta}_0 + x_{i1}\hat{\beta}_1 + x_{i2}\hat{\beta}_2 + x_{i3}\hat{\beta}_3 + x_{i4}\hat{\beta}_4 + x_{i5}\hat{\beta}_5)}}.$$

The dependent values are predicted by $\hat{Y}_i = 1$ when $\hat{p}(\underset{\sim}{x}_i) \geq 0.5$, and $\hat{Y}_i = 0$ when $\hat{p}(\underset{\sim}{x}_i) < 0.5$.

The confusion matrix is a table that is often used to describe the performance of a classification model on a set of predicted data for which the actual data are known following on Table 1. The predicted accuracy is computed by

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}.$$

Table 2.1 : The confusion matrix of actual data (Y_i) and predicted data ($\hat{Y}_i$).

Predicted Data	Actual Data	
	$Y_i = 1$	$Y_i = 0$
$\hat{Y}_i = 1$	True Positive (TP)	False Positive (FP)
$\hat{Y}_i = 0$	False Negative (FN)	True Negative (TN)

2.4 Results and discussion

The estimating coefficient of logistic regression model is obtained from the maximum likelihood (ML), ridge regression (Ridge), lasso, and elastic net methods which transformed to logit model and classified to binary dependent variable. Table 2.1-2.4 present the average percentage various correlation coefficient of predicted accuracy on previous methods. The maximizing percentage are illustrated the performance of these methods.

Table 2.2 : The average percentage of predicted accuracy of maximum likelihood (ML), ridge regression (Ridge), lasso, and elastic net methods on correlation coefficient of 0.1

Distributions	Independent Variables	Sample size	ML	Ridge	lasso	elastic net
Normal	3	n=30	91.88	95.91	**98.72**	97.90
		n=50	94.66	96.86	**99.23**	98.72
		n=100	97.21	97.17	**99.44**	99.15
	5	n=30	89.30	94.82	**98.96**	97.8
		n=50	92.78	95.88	**99.46**	98.71
		n=100	96.16	96.9	**99.57**	99.18
Contaminated Normal (p=0.05)	3	n=30	**91.59**	80.66	83.26	78.10
		n=50	**94.52**	78.73	84.79	78.23
		n=100	**96.98**	79.44	89.68	82.80
	5	n=30	**88.15**	67.81	69.60	66.54
		n=50	**91.57**	66.49	69.02	65.89
		n=100	**94.92**	68.45	70.55	68.26
Contaminated Normal (p=0.1)	3	n=30	**91.47**	78.09	82.34	76.30
		n=50	**94.54**	79.12	87.13	80.26
		n=100	**96.96**	81.26	93.02	87.68
	5	n=30	**87.66**	68.83	72.81	67.72
		n=50	**90.25**	720.6	75.56	71.78
		n=100	**93.47**	75.59	78.47	76.48
t	3	n=30	**98.06**	75.27	91.58	79.28
		n=50	**97.07**	74.65	90.78	79.02
		n=100	**97.03**	92.43	96.44	94.59
	5	n=30	88.55	90.65	**95.34**	92.32
		n=50	92.76	91.39	**96.40**	93.62
		n=100	96.07	91.80	**96.55**	94.09

Table 2.3 : The average percentage of predicted accuracy of maximum likelihood (ML), ridge regression (Ridge), lasso, and elastic net methods on correlation coefficient of 0.5

Distributions	Independent Variables	Sample size	ML	Ridge	lasso	elastic net
Normal	3	n=30	91.90	95.90	**98.73**	97.92
		n=50	94.67	96.84	**99.21**	98.71
		n=100	97.22	97.17	**99.44**	99.16
	5	n=30	89.30	94.82	**98.96**	97.8
		n=50	92.78	95.88	**99.46**	98.71
		n=100	96.16	96.9	**99.57**	99.18
Contaminated Normal (p=0.05)	3	n=30	**91.60**	83.66	83.28	78.11
		n=50	**94.52**	78.72	84.80	78.23
		n=100	**96.97**	79.42	89.68	82.89
	5	n=30	**88.15**	67.81	69.56	66.54
		n=50	**91.59**	66.48	69.02	65.90
		n=100	**94.92**	68.47	70.57	68.28
Contaminated Normal (p=0.1)	3	n=30	**91.48**	78.10	82.34	76.33
		n=50	**94.53**	79.09	87.13	80.26
		n=100	**96.97**	81.26	93.02	87.69
	5	n=30	**87.66**	68.83	72.80	67.72
		n=50	**90.25**	72.06	75.52	71.79
		n=100	**93.40**	75.60	78.49	76.49
t	3	n=30	**98.08**	75.28	91.59	79.29
		n=50	96.70	95.20	**98.51**	97.29
		n=100	**97.07**	92.43	96.44	94.59
	5	n=30	88.56	90.61	**95.34**	92.34
		n=50	92.76	91.38	**96.41**	93.64
		n=100	96.07	91.80	**96.56**	94.11

Table 2.4 : The average percentage of predicted accuracy of maximum likelihood (ML), ridge regression (Ridge), lasso, and elastic net methods on correlation coefficient of 0.99

Distributions	Independent Variables	Sample size	ML	Ridge	lasso	elastic net
Normal	3	n=30	91.76	96.04	**98.65**	97.88
		n=50	94.73	96.98	**99.22**	98.77
		n=100	97.23	97.27	**99.41**	99.16
	5	n=30	89.53	95.22	**99.04**	98.00
		n=50	92.88	96.07	**99.36**	98.62
		n=100	96.14	96.95	**99.54**	99.15
Contaminated Normal (p=0.05)	3	n=30	**91.70**	81.02	83.28	77.92
		n=50	**94.43**	78.87	84.80	77.98
		n=100	**96.99**	80.13	89.82	83.02
	5	n=30	**87.78**	66.68	68.48	65.40
		n=50	**91.52**	65.44	67.86	64.88
		n=100	**94.70**	67.32	69.30	67.03
Contaminated Normal (p=0.1)	3	n=30	**91.46**	78.26	82.31	76.25
		n=50	**94.43**	79.71	87.38	80.57
		n=100	**96.94**	82.19	93.29	88.05
	5	n=30	**87.30**	68.59	72.12	67.18
		n=50	**90.00**	71.71	75.18	71.22
		n=100	**93.60**	74.98	77.96	75.98
t	3	n=30	**97.40**	75.06	91.33	78.96
		n=50	**96.68**	75.14	91.33	78.96
		n=100	97.07	92.74	96.57	**97.88**
	5	n=30	88.58	91.18	**95.36**	92.41
		n=50	92.80	91.78	**96.38**	93.79
		n=100	95.93	92.38	**96.64**	94.28

From Table 2.2-2.4, the lasso is presented a maximum average percentage values for all sample size when the independent variable is simulated from normal distribution. For maximum likelihood method, independent variables via contaminated normal distribution appear the maximum average percentage values. Furthermore, maximum likelihood, lasso, and elastic net when the independent variables are generated from t distribution are shown

the maximum average percentage values in some cases depended on the level of correlation coefficient.

2.5 Application on Real Data

We apply the logistic regression model to analyzed with the breast cancer databases obtained from the University of Wisconsin Hospitals [15]. The binary dependent variables are denoted a symptom of breast cancer as benign and malignant ,and the independent variables are defined by clump thickness (clump), uniformity of cell size (size), uniformity of cell shape (shape) , marginal adhesion (adhesion), single epithelial cell size (epithelial), bare nuclei (bare), bland chromatin (bland), normal nucleoli normal, and mitoses. These data consisted of 100 records. The parameter estimation are approximated from the logistic regression model based on maximum likelihood, ridge, lasso, elastic net method given in Table 2.5-2.6.

Table 2.5 : The percentage of predicted accuracy of maximum likelihood (ML), ridge regression (Ridge), lasso, and elastic net methods on 3 independent variables

Independent Variables	ML	Ridge	lasso	Elastic net
clump,size,shape	44	90	**91**	**91**
clump, size, adhesion	44	89	**93**	92
clump, size, mitosis	44	88	**90**	**90**

Table 2.6 : The percentage of predicted accuracy of maximum likelihood (ML), ridge regression (Ridge), lasso, and elastic net methods on 5 independent variables

Independent Variables	ML	Ridge	lasso	Elastic net
clump,size,shape, adhesion, epithelial	44	92	**93**	**93**
clump,size,shape, bare, bland	44	93	**95**	**95**
clump,size,shape, normal, mitosis	44	92	**93**	**93**

From Table 2.5-2.6, these are apparent that the lasso and elastic net show the maximum percentage of predicted accuracy in most cases except 3 independent variable via clump, size, and adhesion. Therefore, it should be note that lasso and elastic net perform better than the maximum likelihood and ridge since these data correlated between independent variables closed to the normal distribution.

2.6 Conclusions

In this research, we generated correlated independent variables from normal distribution, contaminated normal distribution, and t distribution. The maximum likelihood, ridge regression, lasso, and elastic net are used to estimate parameter on logistic regression model and classified the binary dependent variable. The maximum likelihood method is a good performance when the independent variable is presented on the contaminated normal distribution in most cases. Therefore, the lasso method is a good fit when the independent variables is played on normal distribution and t distribution. For actual data, we are also interested to estimate the class of breast cancer by considering the percentage of predicted accuracy. The results is similar to the simulation data that the lasso and elastic net are a superior over maximum likelihood and ridge regression methods. It is expected because the lasso and elastic net consisted of tuning parameter on the penalty function

which control the interpolating function. If the outlier data occurs on independent variable, the maximum likelihood method will satisfy to solve this problem. On future work, we may focus on Bayesian lasso and Bayesian elastic net to use an expanded hierarchy with conjugate prior for the logistic regression parameter.

Acknowledgements

This Research was financially supported by the Faculty of Science of King Mongkut's Institute of Technology Ladkrabang.

References

[1] AH Lee and MJ Silvapulle. Ridge Estimation in Logistics Regression. *Journal of Communication Statistics-Simulation and computation*. 1988; **17(4)**, 1231-1257.

[2] DE Duffy and TJ Scantner. On a Small Sample Properties of Norm-Restricted Maximum Likelihood Estimators for Logistic Regression Models. *Communication Statistics Theory Mathods*. 1989; **18**, 959-980.

[3] S Cessie and JC Houwellingen. Ridge Estimators in Logistic Regression. *Journal of Applied Statistics*. 1992; **41(1)**, 191-201.

[4] R Tibshirani. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B*. 1996; **58(1)**, 267-288.

[5] PC Bradley and AL Thomas. *Bayesian Methods for Data Analysis*, 3thEd., Chapman & Hall/CRC, London, 2008.

[6] W Gilks, S Richardson and D Speigelhalter. *Markov Chain Monte Carlo in Practice*, Interdisciplinary Statistics, Chapman & Hall, Suffolk, 1996.

[7] S Geman and D Geman. Stochastic Relaxation, Gibbs distributions and the Bayesian Restoration of Images, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1984. Vol. 6, 721-41.

[8] AE Hoerl and RW Kennard. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*. 1970; **12(1)**, 55-67.

[9] D Gamerman. Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference, London: Chapman & Hall, 1997.

[10] A Gelfand, S Hills, A Racine-Poon and A Smith. Illustration of Bayesian Inference in Normal Data models using Gibbs sampling. *Journal of the American Statistical Association*, 1990. Vol. 85, 972-985.

[11] D Lunn, D Speigelhalter, A Thomas and N Best. The BUGS project: Evolution, critique and future directions. *Statistics in Medicine*, 2009. Vol. 28, 3049-3067.

[12] H Zou and T Hastie. Regularization and Variable Selection via the elastic net. *Journal of the Royal Statistical Society. Series B*. 2005; **67(2)**, 301-320.

[13] B Efron, T Hastie, I Johnson and R Tibshirani. Least Angle Regression. *The annals of Statistics*. 2004; **32**, 407-499.

[14] T Hastie, R Tibshirani and J Friedman. *The Elements of Statistical Learning : Data Mining Inference and Prediction*. Springer, 2nd edition, California, 2009, p. 152.

[15] OL Mangasarian and WH Wolberg. *Cancer diagnosis via linear programming*, SIAM News, 1990; **23(5)** , 1-18.