DOI: 10.20472/IAC.2018.935.004

**AUTCHA ARAVEEPORN**

**King Mongkut's Institute of Technology Ladkrabang , Thailand**

# A COMPARISON OF PARAMETER ESTIMATION OF LOGISTIC REGRESSION MODEL BY MAXIMUM LIKELIHOOD, RIDGE REGRESSION, MARKOV CHAIN MONTE CARLO METHODS

## 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.

## Keywords:

Maximum Likelihood, Ridge Regression, Markov Chain Monte Carlo

**JEL Classification:** C13, C15

## 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 (1988) 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 (1989) 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 (1992) 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, 2008). 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, 1996) can approximate the estimator from Gibbs sampling algorithm (Geman, Geman, 1984) based on the posterior distribution.

In this paper, we focus to estimate the coeffifient paremater 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.

## 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

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

where $k$ is a number of independent variable, and $i = 1, 2, \ldots, n$ is a number of observed data. The probability function follows the logistic regression model

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

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

**Method for Estimation**

Maximum likelihood Method

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

$$\mathrm{L}(\beta_0, \beta_1, \ldots, \beta_k) = \prod_{i=1}^{n} p(\underset{\sim}{x}_i)^{Y_i}(1-p(\underset{\sim}{x}_i))^{1-Y_i}. \tag{3}$$

The log likelihood function turns into sum as :

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

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 $\beta_j$, $j = 1, 2, \ldots, k$ by

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

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

Ridge Regression Method

Hoerl, Kennard (1970) 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

$$\hat{\underset{\sim}{\beta}} = (X'X)^{-1} X'\underset{\sim}{Y}, \tag{4}$$

where $\hat{\underset{\sim}{\beta}}$ 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

$$\hat{\underset{\sim}{\beta}}_R = (X'X + \lambda I)^{-1} X'\underset{\sim}{Y}, \; \lambda > 0. \tag{5}$$

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 ($\lambda$).

Markov Chain Monte Carlo Method

The Markov Chain Monte Carlo (MCMC) (Gamerman, 1997) 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, 1990) is an algorithm for MCMC computing. We carry out the WinBUGS Program (Lunn, Spiegelhalter, Thomas, Best, 2009) to obtain the estimating estimator from the posterior distribution function based on MCMC process.

The logistic regression is used the logit model following

$$\text{logit}(p(\underset{\sim}{x}_i)) = \log \frac{p(\underset{\sim}{x}_i)}{1 - p(\underset{\sim}{x}_i)} = \beta_0 + x_{1i}\beta_1 + x_{2i}\beta_2 + ... + x_{ki}\beta_k, \tag{6}$$

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

$$p(\beta_0,\beta_1,...,\beta_k \mid Y_i, \underset{\sim}{x}_i) \propto p(\beta_0,\beta_1,...,\beta_k)\prod_{i=1}^{n} p(\underset{\sim}{x}_i)^{Y_i}(1-p(\underset{\sim}{x}_i))^{1-Y_i}. \quad (7)$$

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

**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. The predicted accuracy is computed by

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

**Table 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) |

**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 2-4 present the average percentage of predicted accuracy on previous methods. The maximizing percentage are illustrated the performance of these methods.

**Table 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 | n=30 | **93.52** | 77.27 | 49.44 |
| | n=50 | **95.65** | 75.85 | 48.92 |

 ISBN 978-80-87927-47-2, IISES

| | | | | |
|---|---|---|---|---|
| (p=0.1) | 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 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 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 2-3, 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 2-3 except n=50 with normal distribution and n=30 with t distribution.

## 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.

## Reference

BRADLEY, P. C. and THOMAS, A. L. (2008). *Bayesian Methods for Data Analysis*, 3th Ed., Chapman & Hall/CRC, London.

CESSIE, S. and HOUWELLINGEN, J. C. (1992) Ridge Estimators in Logistic Regression. *Journal of Applied Statistics*. 1992. Vol. 41, No. 1, 191-201.

DUFFY, D. E. and SCANTNER, T. J. (1989) On a Small Sample Properties of Norm-Restricted Maximum Likelihood Estimators for Logistic Regression Models. *Communication Statistics Theory Mathods.* 1989. Vol. 18, 959-980.

GAMERMAN, D. (1997) Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference, London: Chapman & Hall.

GELFAND, A.; HILLS, S.; RACINE-POON. A and SMITH, A. (1990) Illustration of Bayesian Inference in Normal Data models using Gibbs sampling. *Journal of the American Statistical Association*, 1990. Vol. 85, 972-985.

GILKS, W.; RICHARDSON, S. and SPEIGELHALTER, D. (1996). *Markov Chain Monte Carlo in Practice*, Interdisciplinary Statistics, Chapman & Hall, Suffolk.

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

HOERL, A. E. and KENNARD, R. W. (1970) Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*. 1970. Vol. 12, No. 1, 55-67.

LEE, A. H. and SILVAPULLE, M. J. (1988), Ridge Estimation in Logistics Regression. *Journal of Communication Statistics-Simulation and computation*, 1988. Vol. 17. No. 4, 1231-1257.

LUNN, D.; SPEIGELHALTER, D.; THOMAS, A. and BEST, N. (2009) The BUGS project: Evolution, critique and future directions. *Statistics in Medicine*, 2009. Vol. 28, 3049-3067.