이학석사 학위논문

Comparison of Robustness between Linear Regression and Logistic Regression in Dichotomous Criterion

이분류에서의 선형회귀와 로지스틱 회귀의 비교

2012년 12월

서울대학교 대학원

통계학과

한 지 선

Comparison of Robustness between Linear Regression and Logistic Regression in Dichotomous Criterion

지도교수 김 용 대

이 논문을 이학석사 학위논문으로 제출함.

2012년 12월

서울대학교 대학원

통계학과

한 지 선

한지선의 이학석사 학위논문을 인준함.

2012년 12월

위 원 장 ______________________

부위원장 ______________________

위 원 ______________________

Comparison of Robustness between Linear Regression and Logistic Regression in Dichotomous Criterion

By

Jisun Han

A Thesis
Submitted in fulfillment of the requirements
for the degree of
Master of Science
in Statistics

The Department of Statistics
College of Natural Sciences
Seoul National University
February, 2012

Abstract

As “Big data” has arisen, simplicity and robustness for analysis have been emphasized. Therefore, we focus on the model which is relatively simpler and more robustness against fluctuation of data.

Accurate classification is one of the most important things to decide a model for decision making in practice. Logistic model is known as an accurate model to estimate the probabilities of dependent categories. However, focusing on the goal of classification in practical use, we assume that linear regression can be more efficient in using practical decision making. We start this study to identify the linear regression analysis is better in robustness than the logistic regression analysis.

In simulation, by increasing the number of independent variables, we observe the performances of each method. We try diverse data generated by different models to see the robustness of linear regression analysis. Based on the two conjectures, we analyze the prediction errors of each method. By comparing prediction errors, we conclude the linear regression analysis is the most robust method in our simulation. However, because wee only simulate the equal proportioned two classes, the further study is needed.

Keywords: Linear regression, logistic regression, dichotomous criterion, classification

Student Number: $2011-20254$

Contents

List of Tables

List of Figures

Chapter 1

Introduction

There have been recently increasing studies of 'Big data.' However, it still has some challenges, such as storage, processing, and analysis. For all slightly losing the accuracy of prediction, simpler and more efficient methods are more required as increasing the size of data. From this point of view, classification methods which are mainly used in practice were reconsidered.

Logistic regression is widely used with binary outcome variables. It is one of appropriate methods to develop linear classification models, i.e. models associated with linear boundaries between two groups. There are some strengths of logistic regression.

First, it does not need an additional assumption on the distribution of explanatory data. On the other hand, normality assumption is required in LDA to analyze accurately. In addition, it yields itself to a meaningful interpretation. Coefficients of explanatory variables are related to odds ratios so that analysts can extract the meaning from their results. Last but not least, logistic regression also serves to transform the limited range of a probability, restricted to the range $[0, 1]$, into the real number of range. This makes the transformed values more available to be fitted into a linear function.

However, it is skeptical to work out in the data with a number of independent variables, a kind of large data, too. Besides, it is much considered in scoring because of many strengths we will mention the next chapter, but logistic regression is unknown in classification matter as much as in crediting scores.

Thus, we have started to question whether the simple l_2 linear squared estimation is not really recommendable to analyze dichotomous criterion. Linear regression with l_2 loss can be naively used and less costable because of its simpleness.

Both regressions compare the observed values of the criterion with the predicted values in order to determine if the model that includes the variables more accurately predicts the outcome than the model without a set of variables. There is a reasons why linear regression is not an appropriate method to use in a dichotomous criterion. That is because linear regression violates the measurement assumptions to analyze binary data, but which is fine in classifying binary data. We will mention them the next chapter.

We look through which methods between logistic regression and linear regression are better to classify classes rather than to predict accurately models. This question will help to increase the speed of analysis and to get a solution in 'Big data' because the least squared estimation is simpler and more robust if accuracy of classification is guaranteed.

This paper is organized as follows. In chapter 2, we review linear regression and logistic regression. And, the robustness of linear regression with l_2 loss will be mentioned. In chapter 3, we will look over diverse applications to classification by linear methods and simulate with some conjectures. In chapter 4, the results of simulations will be discussed.

Chapter 2

Reviews of Linear Regression and Logistic Regression

2.1 Logistic Regression for Binary Data

Suppose that x denotes the explanatory variable of interest. The logistic regression model arises from the desire to model the posterior probabilities of the $K \in \mathcal{G}$, a discrete set, classes via linear functions in x, while at the same time ensuring that they sum to one and remain in [0, 1](Friedman et al., 2001). Especially, we only consider the case when $K = 2$ for binary data.

Considering the linear regression model,

$$y_i = \boldsymbol{\beta}^T \mathbf{x}_i + \epsilon \qquad \text{for } i = 1, \ldots, n \tag{2.1}$$

, where $\boldsymbol{\theta} = \{\boldsymbol{\beta} = (\beta_0, \cdots, \beta_p)^T\}$ denote a parameter set, and y_i and $\mathbf{x}_i$ denote ith observations of the dependent and the vector of explanatory variables, respectively.

Then, by using link function, the logistic model for the binary data has the form

$$p(\mathbf{x};\boldsymbol{\theta}) = P(Y=1|\ \mathbf{x};\boldsymbol{\theta}) = \frac{\exp(\boldsymbol{\beta}^T\mathbf{x})}{1+\exp(\boldsymbol{\beta}^T\mathbf{x})}. \tag{2.2}$$

Thus, $P(Y=0|\ \mathbf{x};\boldsymbol{\theta}) = 1 - p(\mathbf{x};\boldsymbol{\theta})$. Using link function, p, some violations of linear regression's assumptions such as normality of ϵ and outrange [0, 1] can be solved.

The *logit transformation* of this model is

$$\ln(\frac{p(\mathbf{x};\boldsymbol{\theta})}{1-p(\mathbf{x};\boldsymbol{\theta})}) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \epsilon. \tag{2.3}$$

Here, the coefficients of a logistic regression model is estimated with MLE method. For a closer information of MLE for logistic regression, please see Friedman et al. (2001).

The advantages of logistic regression are: First of all, scores are interpretable in terms of log odds. Second, Constructed probabilities have chance of making sense(GLU, 2005). However, there are disadvantages of the model such as overinterpretation of some parameters and overfitting.

As considering dichotomous criterion, logistic regression surely performs well in predicting probabilities of scores, not in correcting classification. Besides, the performance of this regression is guaranteed when missing and noised data nonexist. There are however many cases to misspecify models in decision making so that the accuracy is rarely guaranteed. Finally, statistical methods which are widely used in normal situation are skeptical when it comes to 'Big data.' Because of those matters, the simpler method for binary data, simple l_2-norm, also known as the least squared estimation of linear regression gets to an alternative consideration to classify two classes.

2.2 Linear Regression for Binary Data

Linear regression has the form already mentioned as (2.1). Suppose that y_i has 0, or 1 in binary data for simplicity. Because of this change, the conditions, normality of errors and homoscedasticity are not satisfied.

Linear regression has some assumptions which must be satisfied to estimate and make inferences about the coefficients in the analysis. To see the assumptions of linear regression, refer to Menard (2001). However, to analyze binary data, linear regression in dichotomous criterion violates several assumptions such as,

- Linearity

  The conditional mean of the binary criterion given the predictor value must be in the range of [0, 1]. However, this sense is often violated in linear regression so logistic regression is recommendable in binary data.

- Homoscedasticity

  The assumption, "Homoscedasticity", that the variance of the error term, ϵ, is the same or constant for all values of the independent variables must be satisfied to use the OLS method to estimate and make inferences about the coefficients in linear regression analysis. This will be always violated when one has a criterion that is distributed binomially. It means that binary outcome variables don't guarantee homoscedasticity so that linear regression is rarely used in dichotomous criterion.

- Normality

  The errors should be normally distributed in linear regression but this assumption is violated in binary data. That is because the criterion has only two values.

Even though these assumptions should be kept to estimate coefficients, these problems can be disregarded in classification which is the ultimate goal of practical decision making cases. These assumptions are related with estimating and making inference about coefficients of the model. For classifying binary set with large data, robustness of analysis should be regarded. Why we should take into account of robustness in classification will be explained the next section.

2.3 The Definition of Robustness in Classification

Robustness has quite a few meanings in the literature of statistics and machine learning. Especially, we first saw robust classification in Ginodi and Globerson (2011). It was said that being robust to perturbations of the data is usually a desirable property for a classifier because the objective of the learning is to be able to classify new data. In this paper, we also use robustness in the sense of more simplicity and less vulnerability against fluctuations.

One of critical disadvantage of logistic regression is overfitting problem which can be linked to vulnerability. Logistic regression models with large number of independent variables and limited amounts of training data are highly prone to overfitting under maximum likelihood estimation. If the model is overfit, this model is likely to be perturbed by new data. This comes at a cost: it requires much more data to achieve stable results. Thus, logistic regression is not a desirable method in high dimensional data.

Furthermore, in many applications, some of the features x_{ij} may be missing or corrupted in data. Generative models such as LDA(Linear Discriminant

Analysis) typically provide better ways to handle this than logistic regression model. Thus, we will explain robustness of linear regression in detail.

2.4 Robustness of Linear Regression

Recall the l_2 linear regression of dichotomous criterion. We will categorize the strengths of linear regression in robustness into three points.

- When the proportion of two dependent values, 0, and 1 are equal, the performace of the linear regression is the same as LDA's. Classifying rule of the Linear regression is not the same as the LDA rule unless the classes have equal numbers of observations(Ripley 1996, Hastie at al. 2008).

  For reference, let us explain the LR rule and LDA rule. First, let the predictive value

  $$\hat{y} = \hat{f}(\mathbf{x}) = \hat{\boldsymbol{\beta}}^T \mathbf{x}. \tag{2.4}$$

  We fit a linear regression model to each of the columns of $y = (y_1, y_2)$, and the fit given by

  $$\hat{y} = \mathbf{x}(\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T y. \tag{2.5}$$

  As considering minimization of the least squares criterion from (2.1) for binary data,

  $$\Sigma_{i=1}^n (y_i - \beta_0 - \beta^T x_i)^2, \tag{2.6}$$

  then the solution $\hat{\beta} = (\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T y$. Consider the rule: classify to class 2 if $\hat{y}_i > 0.5$ and class 1 otherwise. This rule is the LR rule.

  Second, LDA problem with a binary outcome variable; two classes with class sizes n_1, n_2. Suppose that we model each class density as multivariate Gaussian

$$f_k(x) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} e^{-\frac{1}{2}(x-\mu_k)^T\Sigma^{-1}(x-\mu_k)} \tag{2.7}$$

with a common covariance matrix Σ. In comparing two classes 1 and 2, it is sufficient to look at the log-ratio, and we see that

$$\begin{aligned} \log\frac{P(Y=0|\mathbf{X}=\mathbf{x})}{P(Y=1|\mathbf{X}=\mathbf{x})} &= \log\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} + \log\frac{\pi_1}{\pi_2} & (2.8) \\ &= \log\frac{\pi_1}{\pi_2} - \frac{1}{2}(\mu_1+\mu_2)^T\Sigma^{-1}(\mu_1-\mu_2) \\ &\quad + x^T\Sigma^{-1}(\mu_1-\mu_2), & (2.9) \end{aligned}$$

an equation linear in x.

In practice we do not know the parameters of the Gaussian distributions, and will need to estimate them using our training data. For the estimation, refer to Friedman et al. (2001) in detail. From the (2.8) and estimation, the LDA rule classifies to class 2 if

$$x^T\hat{\Sigma}^{-1}(\hat{\mu}_2-\hat{\mu}_1) > \frac{1}{2}\hat{\mu}_2^T\hat{\Sigma}^{-1}\hat{\mu}_2-\hat{\mu}_1^T\hat{\Sigma}^{-1}\hat{\mu}_1+\log(N_1/N)-\log(N_2/N) \tag{2.10}$$

and class 1 otherwise. Finally, these two rules are not the same when the condition that each classes has equal numbers of observations is not satisfied.

- $E(Y_k|\mathbf{X}=\mathbf{x})$, conditional expectation of each of the random variable, Y_k, seems a sensible goal. However, the predictor, $\hat{f}_k(x)$ of linear regression can be less than 0 or greater than 1, it doesn't seem reasonable(Friedman et al., 2001). But, if we expand linear regression onto basis expansions $h(X)$ of the inputs, this approach gets consistent estimates of the probabilities. Therefore, as the size of the training set n grows bigger, we adaptively include more basis elements so that linear regression onto these basis functions approaches conditional expectations. Thus, even if the proportions in the training set are different, the classification of l_2 regression analysis is reasonable.

- The third one we want to emphasize for robustness of the linear regression is that methods to compute linear regression with large data have been well developed. For example, there is a case, "sampling algorithms for l_2 regression." Drineas et al. (2006) studied sampling algorithms for the basic linear-algebra problem of l_2 regression with large data.

Thus, we assume to take the linear regression analysis in dichotomous criterion with huge data as an alternative method, rather than other methods, LDA, and logistic regression analysis. To make certain of what we assume, we explore earlier studies based on some conjectures and simulate them.

Chapter 3

Applications to Data

3.1 The Earlier Studies

Before starting simulations, we first have two conjectures about when the classification by logistic regression analysis works poorly.

First thing is that performance of logistic regression gets worse even though the true model is logistic when the number of independent variables gets higher. In other words, the analysis of logistic regression works badly in high dimensional data.

Second, we guess that when a model is misspecified, logistic regression performs poorer than linear regression does as increasing the number of explanatory variables. Based on these assumptions, those cases already experimented are inspected.

Based on the conjectures, we study the cases already done. According to Pohlman and Leitner (2003), l_2 linear regression and logistic regression analysis produced very similar results when applied to the same two data sets in their paper. However, estimates the probabilities of dependent category by logistic regression are more accurate, which is commonly known.

By Komarek (2004), logistic regression is reconsidered in data mining and high dimensional classification. As it is written, logistic regression is slow, unstable, and unsuitable for large learning. To solve this problem, logistic regression for fast classification is implemented with large data by regularization to avoid numerical instability, and an efficient implementation.

In other case, animal carcinogenicity experiments are analyzed by the Hoel-Walburg test, and the lifetime incidence test both of which arise as a likelihood score test from a logistic model according to Begg and Lagakos (1990). Comparing Hoel-Walburg test, the lifetime incidence test is inefficient because of only difference of dealing with an important covariate. Lifetime incident test is an example of model misspecification. They study the efficiency of logistic model not robustness of it.

3.2 Simulation Description

Based on the already inspected cases, two conjectures are simulated. We set that two classes have the same proportion. The number of independent variables is written as p.

We set the training set $n_{\mathrm{tr}} = 100$ and test set $n_{\mathrm{ts}} = 10000$. For 100 iterations, we get 100 prediction errors of each model. Even though the number of independent variables get higher, $\|\boldsymbol{\beta}\|_2$ is still fixed to control proper data's dispersion. The explanatory variables are independent with one another.

$$\mathbf{X} = (X_1, \cdots, X_p)^T \sim N(\mathbf{0}_p, c\mathbf{\Sigma}_{\mathrm{I}}), \tag{3.1}$$

where c is constant, and $\mathbf{\Sigma}_{\mathrm{I}}$ is $p \times p$ identity matrix.

By increasing the number of independent variables, we generate the data from 4 cases of true model; logistic model, probit model, LDA model, and QDA model. And, we analyze the data by 4 methods, such as LSE(l_2 linear regression analysis), Logit(logistic regression analysis), Probit(probit analysis), and LDA(linear discriminant analysis).

- The first one is when the true model follows logit model.

$$\text{If } z \sim \text{Logistic}(0,1),\ z + \mathbf{x}^T\boldsymbol{\beta} \sim \text{Logistic}(\mathbf{x}^T\boldsymbol{\beta}, 1) \tag{3.2}$$

  , where $\mathbf{x}$ are observations generated from random variable $\mathbf{X} \sim N(\mathbf{0}_p, c\boldsymbol{\Sigma}_{\text{I}})$. $y = 1$, when $z + \mathbf{x}^T\boldsymbol{\beta} > 0$. Otherwise, $y = 0$.

  From the simulated samples, we analyze the prediction errors of each method. As the number of independent variables(p) gets higher, the difference of accuracy between LSE and Logit gets bigger like Table 3.1. Seeing boxplots in Figure 3.1, the difference can be seen more clearly.

Table 3.1: The Mean of Prediction Error of True: Logistic Model

	p=10	p=20	p=30	p=40
LSE	0.2357	0.2630	0.2880	0.3169
Logit	0.2350	0.2669	0.3034	0.3467
Probit	0.2349	0.2670	0.3033	0.3465
LDA	0.2357	0.2630	0.2880	0.3169

- The second one is when the true model generates from probit model.

$$\text{If } z \sim N(0,1),\ z + \mathbf{x}^T\boldsymbol{\beta} \sim N(\mathbf{x}^T\boldsymbol{\beta}, 1) \tag{3.3}$$

  , where $\mathbf{x}$ are observations generated from random variable $\mathbf{X} \sim N(\mathbf{0}_p, c\boldsymbol{\Sigma}_{\text{I}})$. $y = 1$, when $z + \mathbf{x}^T\boldsymbol{\beta} > 0$. Otherwise, $y = 0$.

When p=10, probit model analysis is the best method for guessing the data generated from probit model. However, after p grows, LSE shows robustness.

Table 3.2: The Mean of Prediction Error of True: Probit Model

	p=10	p=20	p=30	p=40
LSE	0.1630	0.1885	0.2176	0.2412
Logit	0.1603	0.1975	0.2396	0.2634
Probit	0.1597	0.1974	0.2383	0.2644
LDA	0.1630	0.1885	0.2176	0.2412

- The third one is when the true model generates from two different normal distribution with the same covariance matrix, which is for LDA. One class follows $N(\mathbf{0}_p, c\mathbf{\Sigma}_\mathrm{I})$, and the other one is $N(\boldsymbol{\mu}_p, c\mathbf{\Sigma}_\mathrm{I})$, where $\boldsymbol{\mu}_p = (1, 1, 0, \cdots, 0)^T$.

Table 3.3: The Mean of Prediction Error of True: LDA

	p=10	p=20	p=30	p=40
LSE	0.2683	0.2968	0.3185	0.3425
Logit	0.2697	0.3005	0.3340	0.3662
Probit	0.2699	0.3002	0.3343	0.3654
LDA	0.2683	0.2968	0.3185	0.3425

As seen in Table 3.3, LSE and LDA clearly is the best method to guess the data. Because we assign two class with equal proportion, 0.5 and

0.5, their performance, LSE and LDA, is the same in Table 3.3.

- The last one is when the true model generates from two different normal distribution, which is for QDA. One class follows $N(\mathbf{0}_p, c_1\mathbf{\Sigma}_{\mathrm{I}})$, and the other one is $N(\boldsymbol{\mu}_p, c_2\mathbf{\Sigma}_{\mathrm{I}})$, where $\boldsymbol{\mu}_p = (1, 1, 0, \cdots, 0)^T$. In our simulation, we assign $c_1 = 1$, $c_2 = 0.5$ to observe two separated classes properly.

  As seen in Table 3.4, the difference between LSE and Logit gets bigger as the number of explanatory variables increases.

Table 3.4: The Mean of Prediction Error of True: QDA

	p=10	p=20	p=30	p=40
LSE	0.2264	0.2541	0.2769	0.2969
Logit	0.2301	0.2638	0.3007	0.3304
Probit	0.2305	0.2641	0.3005	0.3304
LDA	0.2264	0.2541	0.2769	0.2969

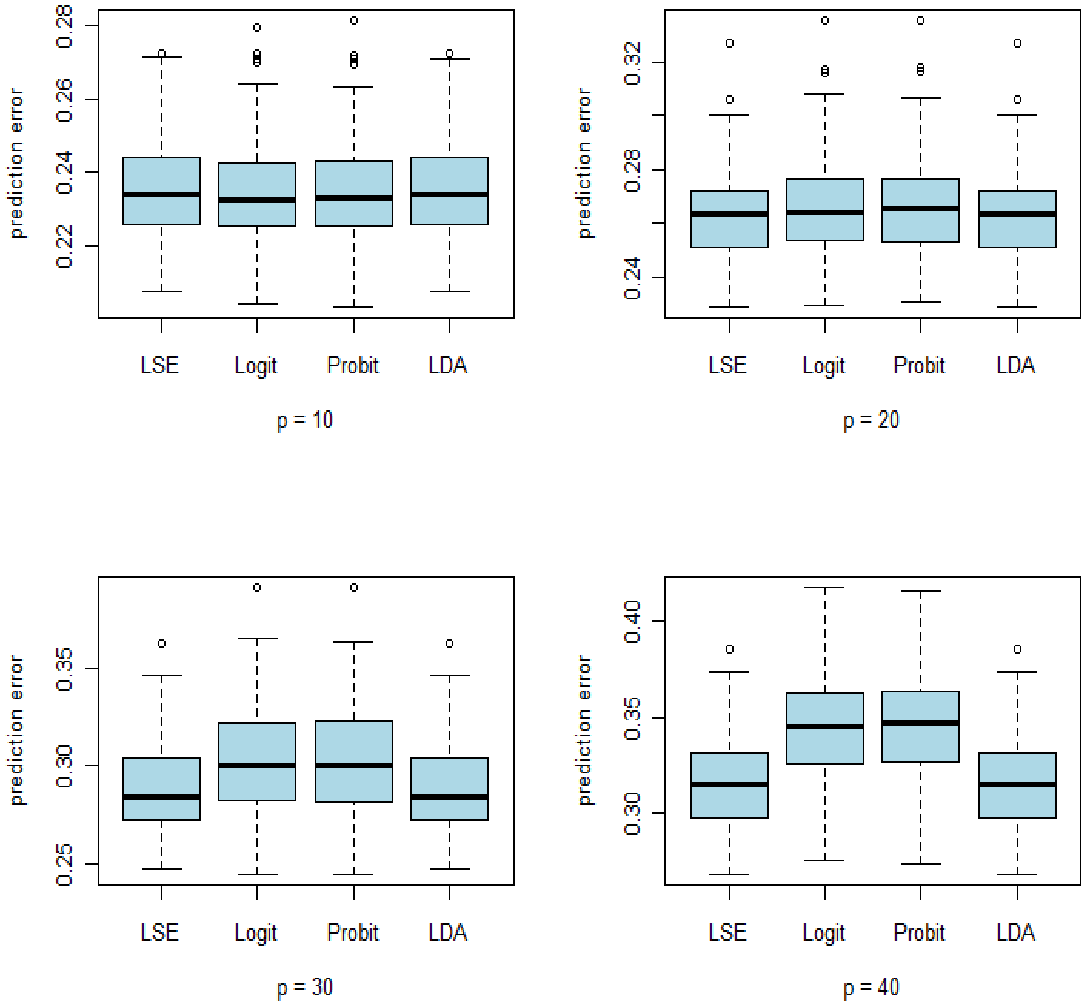

Figure 3.1: Boxplot of Prediction Errors of True: Logistic Model

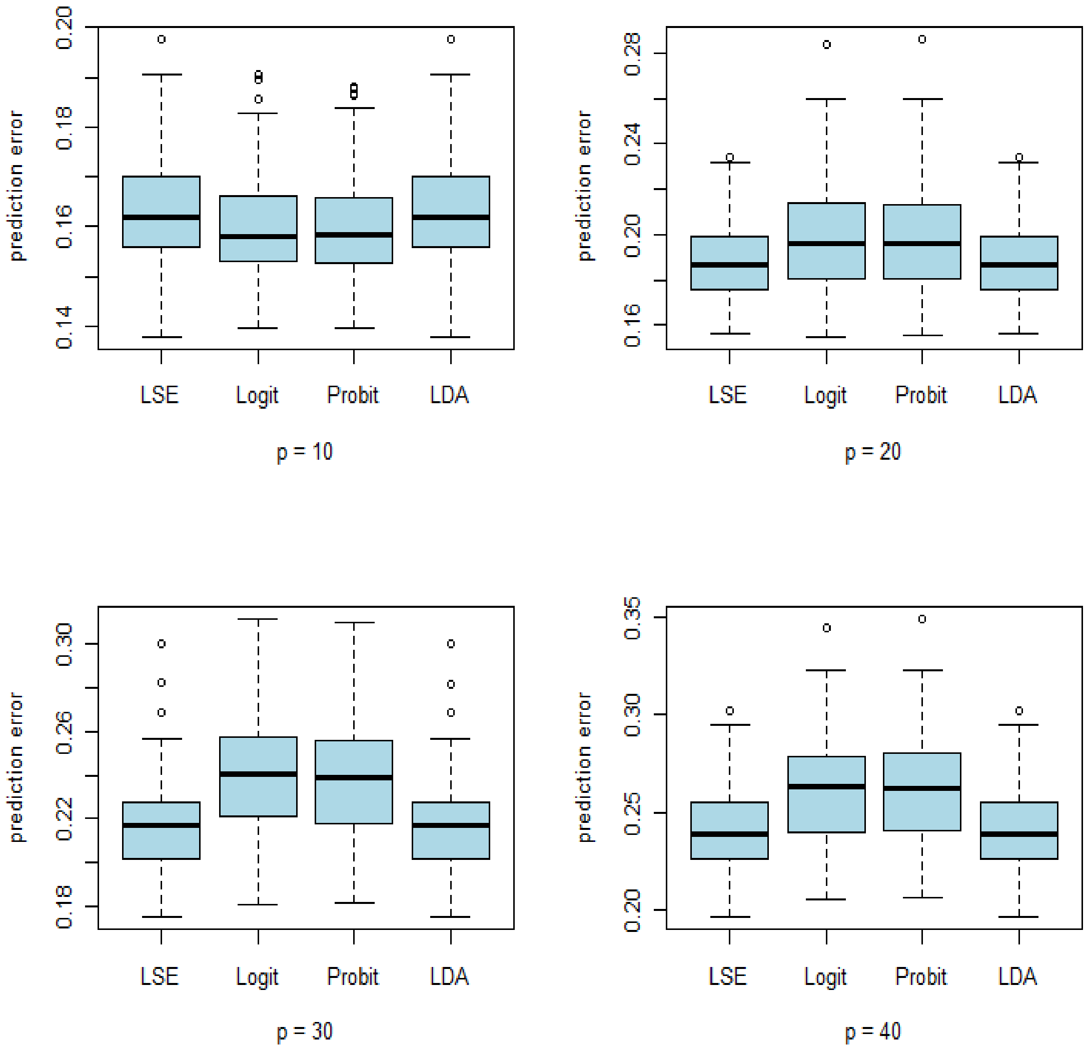

Figure 3.2: Boxplot of Prediction Errors of True: Probit Model

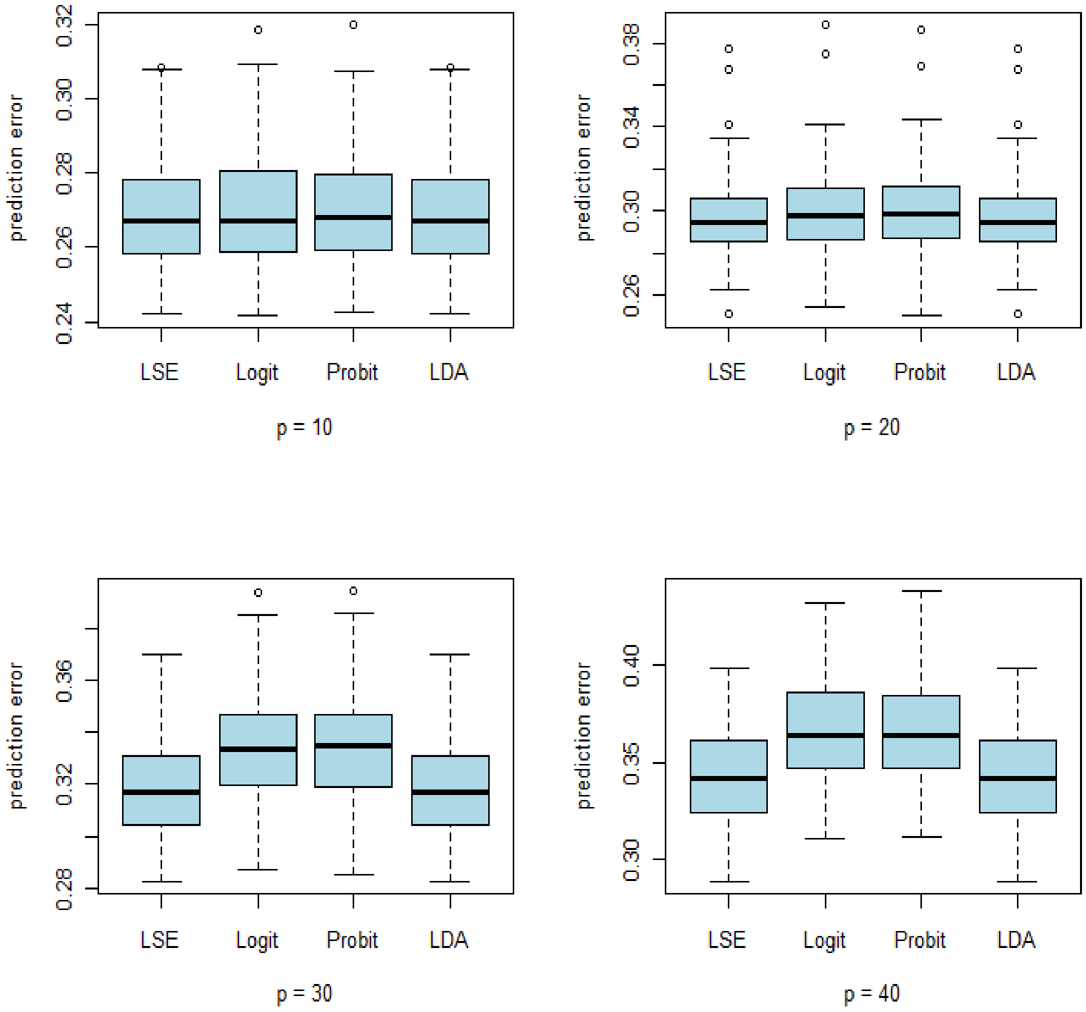

Figure 3.3: Boxplot of Prediction Errors of True: LDA

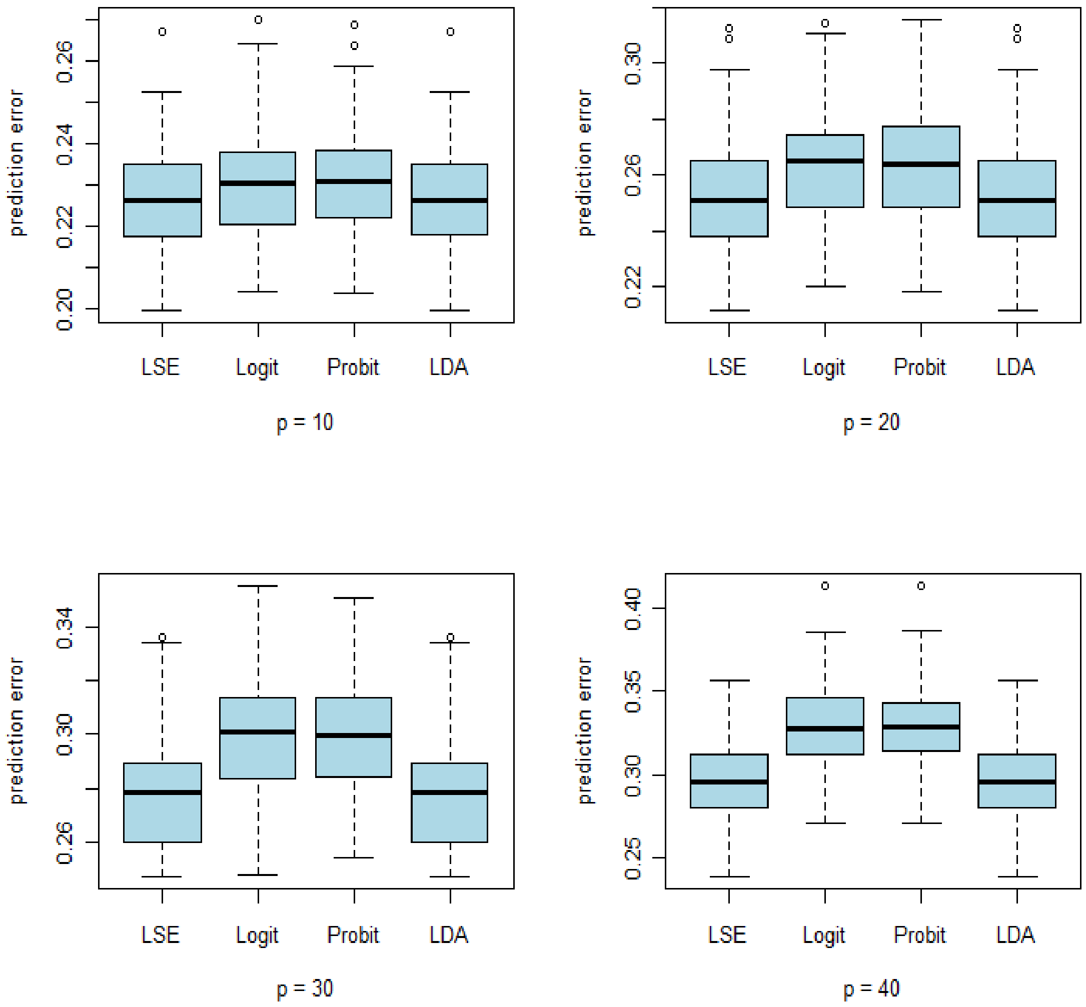

Figure 3.4: Boxplot of Prediction Errors of True: QDA

Chapter 4

Discussion

Among LDA, probit analysis, logistic analysis, and l_2 linear regression analysis, l_2 linear regression analysis is the best method for classification with high explanatory variables in our simulations. By simulation, LSE is a recommended method for dichotomous criterion with the same proportion. As p grows, the performance of LSE is surely better than the one of logistic regression analysis. Robustness of LSE is explained.

We need a further study to identify the performance of LSE in a case two classes are not equally distributed. Simulations with correlated independent variables are done but, they are abbreviated. That is because their performance seems quite the same with the simulations we have described above although the accuracy of prediction of it decreases.

Bibliography

Begg, M.D. and Lagakos, S. (1990). On the consequences of model misspecification in logistic regression. *National Institute of Environmental Health Science*, **87**, 69.

Drineas, P. and Mahoney, M.W. and Muthukrishnan, S. (2006). Sampling algorithms for l 2 regression and applications. *Proceedings of the seventeenth annual ACM-SIAM symposium on Discrete algorithm*, 1127–1136.

Friedman, J. and Hastie, T. and Tibshirani, R. (2001). *The elements of statistical learning*, **1**, 101–109, 135–136.

Ginodi, I. and Globerson, A. (2011). Gaussian Robust Classification. *arXiv preprint arXiv:1104.0235*, 8–9.

GLU, A. (2005). CREDIT SCORING METHODS AND ACCURACY RATIO. 33–35.

Komarek, P. (2004) Logistic regression for data mining and high-dimensional classification *Robotics Institute*, 222.

Menard, S. (2001) *Applied logistic regression analysis*, **106**, 4–5.

Pohlman, J.T. and Leitner, D.W.(2003). A comparison of ordinary least squares and logistic regression.

Ripley, B.D. (1996). Pattern recognition and neural networks. *Cambridge Uni. Press, Cambridge.*

국문초록

데이터가 대두되며 분석에 대한 simplicity와 robustness가 강조되고 있기 때문에 간단하면서 상대적으로 데이터 변화에 로버스트한 모델을 찾는 것이 이 연구의 관심사였다.

실생활에서 의사결정 시 모델에 가장 중요한 것 중 하나는 데이터를 잘 분류하는 것이다. 로지스틱 모델이 반응 변수에 대한 probability를 잘 추정하여 score test의 정확성을 인정받았다. 하지만 반응 변수에 대한 값을 잘 예측하는 것보다 결국 분류의 목적에 착안하여 선형 회귀 모델을 효율성 있게 사용할 수 있지 않느냐는 의문을 가지게 되었다. 결국 데이터 분류를 목적으로 하는 자료에서 로지스틱 모델이 이 분류뿐만이 아니라 일반적으로 많이 쓰이는 선형회귀분석보다 더 나은 지에 대한 것을 검증하고자 이 연구를 시작하였다.

로지스틱 모델이 새로운 데이터에 대한 로버스트가 상대적으로 LDA보다 약하다는 것이 연구가 되어 있었고(GLU, 2005), Overfitting에 대한 문제 또한 있다는 것이 알려져 있었다. 또한, 이분류에서 두 클래스의 비율이 같지 않을 경우, LDA와 선형회귀의 performance가 동일하지 않다(Friedman et al., 2001).

시뮬레이션을 시행하기 전에 두 클래스의 비율은 같다고 설정하였고, 독립 변수가 서로 독립인 경우와 독립이지 않은 경우, 두 가지를 모두 살펴보았다. 이 조건 아래, 추론한 두 가지 가설을 검증하기 위해 시뮬레이션

을 하였다. 첫 번째는 독립 변수에 대한 모델이 로지스틱 분포를 따른다고 하더라도 독립 변수의 수가 증가할수록 로지스틱 모델의 정확성이 선형회귀보다 현저하게 떨어진다는 것이다. 두 번째로는, 독립 변수의 모델이 misspecification되어 있을 때 로지스틱 모델의 정확성이 떨어진다는 것이다.

이 두 가지 가설들을 바탕으로 시뮬레이션을 해 본 결과, 독립 변수의 수가 증가할수록 로지스틱 모델의 예측 에러가 급격하게 증가하는 것을 볼 수 있었다. 이 시뮬레이션 검증을 통해서 독립 변수의 수가 증가할 수록 이 분류에서 로지스틱 분석을 사용하는 것을 재고하고 선형회귀를 사용해도 무방하다는 것을 알 수 있었다. model misspecification에 관계없이 선형회귀분석이 독립변수가 증가함에 따라 이분류에서는 월등하게 로지스틱 모델보다 더 좋은 것으로 나타났다. 각 모델에 대한 비교는 예측 에러(prediction error)로 하였다.

하지만, 두 클래스의 비율이 같을 때만을 시뮬레이션 해보았고, noise나 missing에 대한 robustness of linear regression analysis 검증이 더 필요하다.

주요어 : 로지스틱 회귀분석, 선형회귀분석, 이분류

학 번 : 2011 – 20254

감사의 글

석사 2학년의 기간이 짧다고 하면 짧고 길다고 하면 긴 기간이었는데 그 동안 좋은 분들, 많은 배움 얻어 간 것 같아 이 기회를 빌어 저를 항상 도와주셨던 많은 분들께 감사드립니다.

먼저 저에게 무한한 사랑과 관심을 주시고 또 도움을 주셨던 김용대 교수님께 정말 감사드립니다. 앞으로도 자주 찾아뵙도록 노력하겠습니다. 많은 가르침 외에도 즐거움 또한 아낌없이 주셨던 것 감사드립니다. 전종우 교수님, 박병욱 교수님 이하 모든 통계학과 교수님들 진심으로 존경합니다.

언제나 북적북적했던 우리 연구실 가족들께도 많은 감사드립니다. 항상 웃음이 끊이질 않는 토니안, 종준 오빠 그리고 유미 언니 많이 챙겨주셔서 감사드립니다. 항상 세심하게 챙겨주시고 조언 아끼지 않았던 상인 오빠, 힘들 때 ”화이팅!” 문자 보내주시던 재석 오빠 감사드립니다. 그리고 묵묵히 저의 뒷자리를 지켜주셨던 병엽 오빠 결혼 축하드립니다. 종종 맛있는 과일도 챙겨주고, 취업 준비에 큰 도움 주셨던 미애 언니 종종 뵈었으면 좋겠습니다.

연구실 권상우, 권성훈 박사님, 자상하신 광수 박사님, '자상함'의 대명사 주유 오빠, 항상 따뜻한 편지로 감동시키는 상미 언니, 듬직한 우성 오빠, 놀리고 괴롭혀도 공부 가르쳐 줄 때는 친절한 민우 오빠, 미국에서 곧 돌아올 승환 오빠, 젠틀맨 원준 오빠, 자타 공인 슈퍼맨 재성 오빠, 연구실 생활 잘해서 대견스러운 우리 슬기, 말 잘 듣는 구환이, 세민이, 동하 모두 감사합니다.

마지막으로 우리 가족들, 엄마, 아빠, 지수 그리고 막내 지원이. 모두 건강하고 행복하길 바랍니다.

대학원을 진학하여 2년이라는 시간 동안, 아쉬움이 많았지만 많은 것을 배울 수 있는 시간이었습니다. 그것들은 제가 사회에 진출했을 때 탄탄한 자산이 될 것입니다. 제 주변에 계신 모든 분들께 감사의 인사드립니다. 모두

건강하세요.

2012년 12월 한지선