11-6-2015

Analysis of Rheumatoid Arthritis Data using Logistic Regression and Penalized Approach

Wei Chen
University of South Florida, weichen1@mail.usf.edu

Analysis of Rheumatoid Arthritis Data using Logistic Regression and Penalized Approach

by

Wei Chen

A thesis submitted in partial fulfillment
of the requirements for the degree of
Master of Arts
Department of Statistics
College of Arts and Sciences
University of South Florida

Major Professor: Dan Shen, Ph.D.
Chris Tsokos, Ph.D.
Yuncheng You, Ph.D.

Date of Approval:
October 28, 2015

Keywords: logistic regression, shrinkage method, rheumatoid arthritis clinical trial data

DEDICATION

This dissertation is dedicated to my parents, who have supported me during these years.

ACKNOWLEDGMENTS

I would like to thank my advisors Dr. Dan Shen for his continuous support and guidance during this process. I also would like to thank my parents for their support during these years.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

In this paper, a rheumatoid arthritis (RA) medicine clinical dataset with an ordinal response is selected to study this new medicine. In the dataset, there are four features, sex, age, treatment, and preliminary. Sex is a binary categorical variable with 1 indicates male, and 0indicates female. Age is the numerical age of the patients. And treatment is a binary categorical variable with 1 indicates has RA, and 0 indicates does not have RA. And preliminary is a five class categorical variable indicates the patient's RA severity status before taking the medication. The response Y is 5 class ordinal variable shows the severity of patient's RA severity after taking the medication.

The primary aim of this study is to determine what factors play a significant role in determine the response after taking the medicine. First, cumulative logistic regression is applied to the dataset to examine the effect of various factors on ordinal response. Secondly, the ordinal response is categorized into two classes. Then logistic regression is conducted to the RA dataset to see if the variable selection would be different. Moreover, the shrinkage methods, elastic net and lasso are used to make a variable selection on the RA dataset of two-class response for the purpose of adding penalization to increase the model's robustness.

The four model results were compared at the end of the paper. From the comparison result, logistic regression has a better performance on variable selection than the other three approaches based on P-value.

CHAPTER 1: INTRODUCTION

1.1 Background

Rheumatoid arthritis (RA) is a type of autoimmune arthritis which is caused by the faulty immune system and creates inflammation in joints [1]. Moreover, it may cause the tissue that lines the inside of joints to thicken that lead to swelling and pain in the joints. The elastic tissue "cartilage" which is the cover of the end edges of bones in joints may be damaged if the disorder of inflammation is not controlled. The joint spacing between bones will become smaller due to loss of cartilage. Joints will be unstable and painful. Joints may lose their mobility when the condition is severe. Joint damage cannot be reversed and may lead to deformity. Doctors recommend early diagnosis and treatment to control rheumatoid arthritis since joint damage can occur early. Joints of the hands, feet, wrists, elbows, knees and ankles can be severely effected by RA. The joint effect is usually symmetrical which means the right knee joint will be effected if the left one is affected. Because RA also can influence the entire body system. That is why it is considered as a systemic disease [2].

According to the national statistics, There are about 1.5 million people in the United States have rheumatoid arthritis (RA). The amount of female RA patients is three times as the amount of male RA patients. For female, RA most begins at 30 to 60. In contrast, it often begins later in life for male. The odds of having RA will be increasing if there are RA patients in families. But the most RA patients do not family history of the disease [2].

The cause of rheumatoid arthritis is not clear exactly. The abnormal response of the immune system is considered as the main reason for joint inflammation and. The risk factors include gene, infectious agents, female hormones, obesity, environment and the body's response to stressful events [2].

Rheumatoid arthritis is a chronic immune disease that cannot be cured. But some medications can slow disease activity. Corticosteroids which the major medications are prednisone, prednisolone and methylprednisolone have an obvious effect of anti-inflammatory. They are used to control the potential damage of joint inflammation. DMARD is an abbreviated form of the disease-modifying anti-rheumatic drug which is used to improve the course of RA. Biologics may slow, improve and stop RA when other treatments do not help for some patients. Every biologic may block a specific step in RA progress. Abatacept, adalimumab, anakinra, certolizumab pegol, etanercept, infliximab, golimumab and rituximab are the most popular biologics [2].

1.2 Study Objectives

This research aims to apply the methodologies to the rheumatoid arthritis clinic data to do the significant factor analysis.

The objectives of this research are to (1) conduct a cumulative logistic regression to examine the influence of various factors in rheumatoid arthritis clinical trial data, then we divided the response variable into two classes and apply a logistic model to make the variable selection (2) apply penalized approaches to the RA data of the two-class response variable to determine which factors affect the outcome significantly and compare the two shrinkage methods according to the results.

1.3 Organization

This document is organized as follows:

- Chapter 2 is a data introduction. Detailed description of the variables are presented. The plots of relationship among the variables are shown.
- Chapter 3 presents the methodology applied, including model formulation and model fitting.
- Chapter 4 presents the results for each model described in the methodology section.
- Chapter 5 includes conclusion of the study.

CHAPTER 2: DATA

The rheumatoid arthritis clinical trial (RA) dataset is used for this analysis. The dataset is extracted from “Statistics in Medicine, 1994” which is a peer-reviewed medical statistics journal published by Wiley. Established in 1982 [9]. The data is a record of the clinical trial of 302 RA patients that are measured on a five-level response which is the rheumatoid arthritis self-assessment score from 1 to 5. There are 302 observations on 5 variables in the dataset. The following refers to the description of the RA data.

y	1 for very good after 3 months clinical trail 2 for good after 3 months clinical trail 3 for fair after 3 months clinical trail 4 for poor after 3 months clinical trail 5 for very poor after 3 months clinical trail
baseline	1 for very good before clinical trail 2 for good before clinical trail 3 for fair before clinical trail 4 for poor before clinical trail 5 for very poor before clinical trail
sex	0 for male 1 for female
age	Patients' age recorded at the baseline
trt	0 for the drug group 1 for the placebo group.

In Rheumatoid Arthritis clinical trial dataset, “age” is a continuous variable. It describes the patients’ age in clinical trial. “y” and “baseline” are ordinal variables. “y” and “baseline” refer to the patients’ self-assessment score of rheumatoid arthritis after three-month medication

treatment and original self-assessment score of rheumatoid arthritis when they enrolled in this study. "sex" is a binary variable which is the patients' gender. The patients are divided by two groups to test the efficacy of the drug. One group took the new medication for an experimental treatment. The other group which is the non-drug therapy group is recorded as the placebo group in the study.

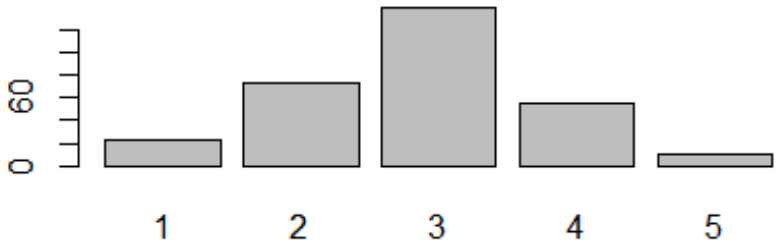


Figure 2.1. RA self-assessment score before the medication treatment

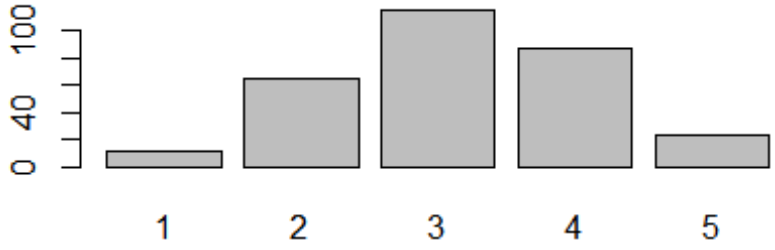


Figure 2.2. RA self-assessment score after the medication treatment

From Figure 2.1 and Figure 2.2, we can see the population change of different RA score during the clinical trial.

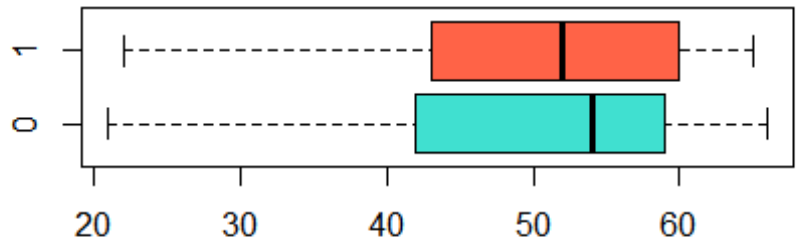


Figure 2.3. Age distribution of male and female RA patients

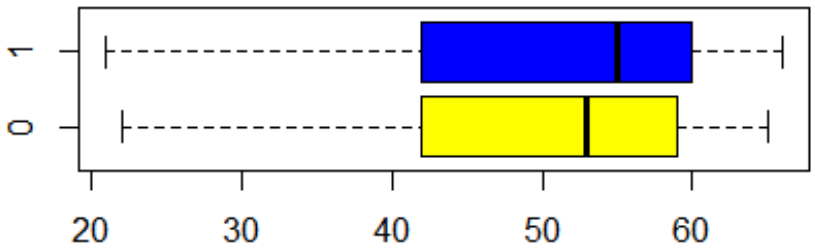


Figure 2.4. Age distribution of drug group and non-drug group

Figure 2.3 shows the RA male and female patients' age. In Figure 2.4, age distribution of RA patients in drug group and non-drug is shown. Most of patients in this clinical trial are at age of 40 to 60.

CHAPTER 3: METHODOLOGY

This chapter presents the methodologies applied to achieve the goals of this study. The methodologies of logistic regression, ordinal logistic regression, elastic net regression and lasso regression are shown in detail in this chapter.

3.1 Logistic Regression

Logistic regression is used to analyze the relationship between the dependent variables which are categorical [3]. It is called a cumulative or ordinal logistic regression when used to predict the probabilities of ordinal outputs.

3.1.1 Model Formulation

Suppose we have random variables (X, Y), where X ∈ R, Yi is a dichotomous variable and Y∈ {0, 1}. Y is defined as a Bernoulli random variable. So the success probability is Pr(Y =1 |X) = p(X), where p(X) is the logistic function. That is

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}} \quad (1)$$

This formula generates from $p(X) = \beta_0 + \beta_1 X$ which is the probability of simple linear regression.

From equation (1) we have:

$$\frac{p(X)}{1-P(X)} = e^{\beta_0+\beta_1} \quad (2)$$

Manipulate the equation (2) by taking the logarithm of both sides, we have:

$$\log(\frac{p(X)}{1-P(X)}) = \beta_0 + \beta_1 X \quad (3)$$

Where $e^{\beta_0+\beta_1} = \frac{p(X)}{1-P(X)} \in (0, \infty)$ is the odds and $\log(\frac{p(X)}{1-P(X)})$ is the log odds.

Then we can derivate the odds ratio formula from equation (3), that is

$$\text{OR}= \frac{\frac{p(X+1)}{1-P(X+1)}}{\frac{p(X)}{1-p(X)}} = \frac{e^{\beta_0+\beta_1(X+1)}}{e^{\beta_0+\beta_1 X}} = e^{\beta_1}$$

The logistic model is called a cumulative or ordinal logistic model when the depend variable is ordinal. Let the ordinal response variable be Y=1, 2, 3...j [5]. The relevant probabilities are $\{\pi_1 + \pi_2 + \cdots \pi_j\}$, the corresponding cumulative probability of a response less than equal to j is:

$$p(Y \le j) = \pi_1 + \pi_2 + \cdots \pi_j = \frac{\exp(\alpha_j + \beta x)}{1+\exp(\alpha_j + \beta x)}$$

The cumulative logit is defined as

$$\log\left(\frac{p(Y \le j)}{p(Y > j)}\right) = \log\left(\frac{p(Y \le j)}{1-p(Y \le j)}\right) = \log\left(\frac{\pi_1 + \cdots + \pi_j}{\pi_{j+1} + \cdots + \pi_J}\right)$$

3.1.2 Model Fitting

Typically, we use the ordinary least squares (OLS) approach by estimating the coefficients β_0 and β_1 in linear regression. However, maximum likelihood is a better method compared with non-linear least squares to fit a logistic regression model [3]. The basic intuition is finding the estimates of β_0 and β_1 for the predicted probability of p(X) to calculate a value close to 1 for all

the success and a value close to 0 for all the failure. The maximum likelihood function is:

$$l(\beta_0, \beta_1) = \prod_{i:y_i=1} p(x_i) \prod_{i':y_i'=0} (1 - p(x_i'))$$

We maximize this likelihood function by choosing the estimates of β_0 and β_1.

Deviance is an appropriate approach to test the fit of a dataset in a logistic regression model. This approach is similar to the sum of squares calculation in linear regression in some ways. The model fits well if its deviance is small. The formula is defined as:

$$Dev(V) = -2\left[\log\left(l(V)\right) - \log\left(l(V_s)\right)\right]$$

Where V_s is the saturated model which means a model with the perfect fit. V is the fitted model. $\log\left(l(V)\right)$ and $\log\left(l(V_s)\right)$ are the log-likelihood of V and V_s. $Dev(V)$ follows chi-square distribution.

3.2 Shrinkage Methods

Shrinkage method refers to a useful approach of estimation or prediction when two or more of the independent variables in a regression model are correlated. In other words, it is a method of fitting a regression model with all predictors [4]. The goal of this method is reducing the related variance significantly by regularizing or shrinking the coefficient estimates towards zero to make the variable selection. In this chapter, we introduce two shrinkage regression models which are elastic net and lasso.

3.2.1 Model Formulation

Suppose we have a multiple regression model.

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

Where $Y = (y_1, \cdots, y_n), n \times 1$ and $X = (x_1, \cdots, x_p), n \times p$

The formula of ordinary least squares estimates of parameter β is:

$$\beta = (x^T x)^{-1} x^T y$$

Where $(x^T x)^{-1}$ is inverse of matrix of $x^T x$ and $x^T y$ is the vector of their sums of products with y. The ordinary least square estimator β is the best fitting and best linear unbiased estimator. Its variance is:

$$\text{var}(\beta) = (x^T x)^{-1} \sigma^2$$

Where the matrix $(x^T x)^{-1}$ is near singular, so $\text{var}(\beta)$ will have many elements. β is not clearly stated and explained under exact collinearity. In such a situation, shrinkage methods can be used to trade bias for variance [4].

Elastic net is one of regularization models. It prevents coefficients of linear regression models with many correlated variables with high variance. It can shrink the coefficients of correlated predictors towards zero to make a variable selection [5].

Elastic net regression is based on the least squares coefficient estimates of linear regression that is using the values that minimize residual sum of squares, which is

$$RSS = \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2$$

The values of coefficient estimates of elastic net minimize $RSS + \lambda \sum_{j=1}^{p} \beta_j^2 + \lambda \sum_{j=1}^{p} |\beta_j|$ which the formula is

$$RSS + \lambda \sum_{j=1}^{p} \beta_j^2 + \lambda \sum_{j=1}^{p} |\beta_j| = \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p} \beta_j^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

$$\min\left\{\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2\right\}$$

$$\text{subject to } \sum_{j=1}^{p}\beta_j^2 \le t_2$$

$$\sum_{j=1}^{p}\left|\beta_j\right| \le t_1$$

Where $\lambda \ge 0$ is the tuning parameter and $\lambda\sum_{j=1}^{p}\beta_j^2+\lambda\sum_{j=1}^{p}\left|\beta_j\right|$ is the shrinkage penalty. t_1 and t_2 refer to the tuning parameters of lasso and ridge.

The Lasso is a relatively alternative method of elastic net. It also can make the variable selection [5].

$$RSS+\lambda\sum_{j=1}^{p}\left|\beta_j\right| = \sum_{i=1}^{n}(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij})^2+\lambda\sum_{j=1}^{p}\left|\beta_j\right|$$

$$\min\left\{\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2\right\}$$

$$\text{subject to } \lambda\sum_{j=1}^{p}\left|\beta_j\right| \le t_1$$

We see that the lasso and elastic net regression have similar formulations. The difference is the penalties. The lasso uses an L1 penalty. The L1 penalty has the effect of forcing some of the coefficient estimates to be exactly equal to zero when the tuning parameter λ is sufficiently large.

3.2.2 Model Fitting

λ is defined as the tuning parameter. The value of λ is greater than zero. The usage of tuning parameter λ is controlling the coefficient estimates. The following is the explanation of the effects of different values of λ.

(1) When λ = 0, the least squares estimates will be produced. It means there is the effect

for penalty term.

(2) When $\lambda \to \infty$, the coefficient estimates will be close to 0. It indicates the growth of the effect of shrinkage penalty.

Cross-validation is the method of selection of tuning parameter. The primary concept is choosing a grid of λ values and computing the cross-validation error for each value of λ. Then we select the best tuning parameter which has the smallest cross-validation error. At last we fit the model by using all observations and the best tuning parameter**.** The function is given by

$$CV_{(k)} = \frac{1}{k}\sum_{i=1}^{k} MSE_i$$

Where MSE_i is the mean square error.

$$CV_{(k)} = \frac{1}{k}\sum_{i=1}^{k} Err_i$$

Where $Err_i = I(y_i \neq y_i)$**.**This formula is defined as cross-validation on classification problems.

CHAPTER 4: RESULTS

This chapter presents the results by using different models. The results that analyzed by logistic regression are illustrated in the first part. Two Logistic regression models have been applied to the RA dataset to make a comparison on variable selection for different response classifications. A comparison between two regularization models on variable selection in the RA dataset when response variable has been divided into two classes is presented in 4.2.

4.1 Logistic Regression

The logistic regression model is applied to analyze the RA dataset in the first part. First, we aim at the response variable "y" which is the final self-assessment score of rheumatoid arthritis to evaluate the impact of the other variables at different self-assessment scores. The predictor variables are the age of patients, the gender of patients, the self-assessment score of rheumatoid arthritis at each baseline and the different treatment groups of patients.

Table 4.1. Ordered log-odds, odds ratio, p-value and 95% CI of predictor variables

Variable	*Ordered log odds*	*Odds ratio*	*P-value*	*95% CI*
sex	*-0.15788*	*0.8539552*	*0.2458468*	*[0.6539397, 1.1148591]*
trt	*-0.48097*	*0.6181837*	*0.00010168*	*[0.4847353, 0.7874804]*
baseline2	*0.63288*	*1.8830283*	*0.01754823*	*[1.1174971, 3.1800030]*
baseline3	*1.13816*	*3.1210301*	*0.000008*	*[1.8958975, 5.1562383]*
baseline4	*2.46945*	*11.8159101*	*1.4023E-17*	*[6.7195089, 20.9077778]*
baseline5	*4.0324*	*56.3961248*	*1.91E-21*	*[24.7443444, 130.8564469]*
age	*-0.01229*	*0.9877886*	*0.02941853*	*[0.9768979, 0.9987502]*

From the column of ordered log-odds in Table 4.1, we could see that the self-assessment score of rheumatoid arthritis “y” is expected to change since the predictor variables increase by one unit or level. For the dichotomous variable “sex”, the ordered log-odds of females was -0.15788 less than males in a higher self-assessment score when the other variables in the model are held constant. The variable “trt” which is defined as dichotomous treatment group “drug” and “placebo” has an ordered log-odds of -0.48097. This indicates the decreasing in ordered log-odds of comparing placebo group to drug group. With the increase in “age” at higher self-assessment score, the ordered log-odds is decreasing by 0.01229. For the ordinal predictor variable “baseline”, every ordered logit odds of different level of “baseline” is increasing with the higher self-assessment score of rheumatoid arthritis. The values of odds ratio of “age”, “sex” and “trt” are less than 1. That indicates the exposure associated with lower odds of outcome. In contrast, each score of the “baseline” has a greater magnitude for the response variable “y” compared to other three predictor variables. From all results of p-value, we could summarize that the variable “trt” is statistically highly significant as $P < 0.001$. The “baseline” has a great effect for all responses of “y” when its score values are 2, 3, 4 and 5. The variables “age” and “sex” do not influence final self-assessment score of rheumatoid arthritis “y” significantly. The range of 95% CI for all predictor variable are shown in the column. All 95% CI of all variables do not cross 0. It indicates that the parameter estimates are statistically significant.

Then we divide the response variable “self-assessment score of rheumatoid arthritis “y” for two groups “good for $y < 3$” and “severe for $y >= 3$” in order to make a comparison between the significate variable selection of the logistic regression on different classification of response.

Table 4.2. Two-class log-odds, odds ratio, p-value and 95% CI of predictor variables

Variable	*log odds*	*Odds ratio*	*P-value*	*95% CI*
sex	*0.13809*	*1.148085*	*0.2458468*	*[-0.4865685, 0.7627599]*
trt	*-0.327838*	*0.7204797*	*0.2392319037*	*[-0.8738030, 0.2181268]*
baseline2	*1.09500*	*2.989191*	*0.0270104426*	*[0.1244872, 2.065518]*
baseline3	*1.77808*	*5.918502*	*0.0002086449*	*[0. 8383084, 2.717858]*
baseline4	*2.10396*	*8.198575*	*0.0001966691*	*[0.9964122, 3.211508]*
baseline5	*16.97979*	*0.00000023*	*0.9812004129*	*[-0.00139533, 1429.297]*
age	*-0.002174*	*0.9877886*	*0.6648019462*	*[0.02692466, 0.02257533]*

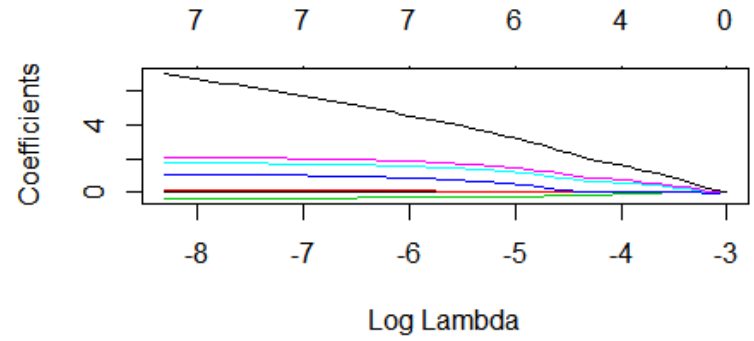
In Table 4.2, the values log-odds and odd ratio are changed compare to the values in Table 4.1 due to different classification of response variable ."baseline" has the least p-value. It indicates "baseline" is the most significate variable in the model when the value of "baseline" are 2, 3, and 4. "sex", "trt" and "age" do not have the influence on final RA score. The values of 95% CI of variables "sex" "trt" and "baseline5" cross zero. This indicates the parameter estimates are not statistically significant.

From the results of Table 4.1 and Table 4.2, we can conclude that the variable selections are different which we the classification of response are different. The most significant difference is that "trt" no longer influence the response variable "y" when "y" is dichotomous.

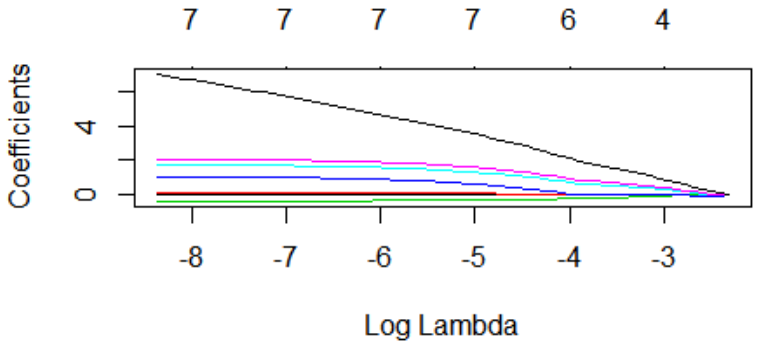
4.2 Shrinkage Methods Comparison

In this section, we apply lasso and elastic net regression models to the RA dataset to estimate the performance of the variables. We divide the response variable "self-assessment score of rheumatoid arthritis "y" for two groups "good for y < 3" and "severe for y>= 3" Then we compare the results of these two shrinkage methods. These two methods would shrink some coefficient estimates to 0. The predictor variables which have the coefficient estimates of zero are

not considered as the significant factors [8]. The plots in Figure 4.2 illustrate the relationship between regression coefficients and the penalty parameters in lasso and elastic net.



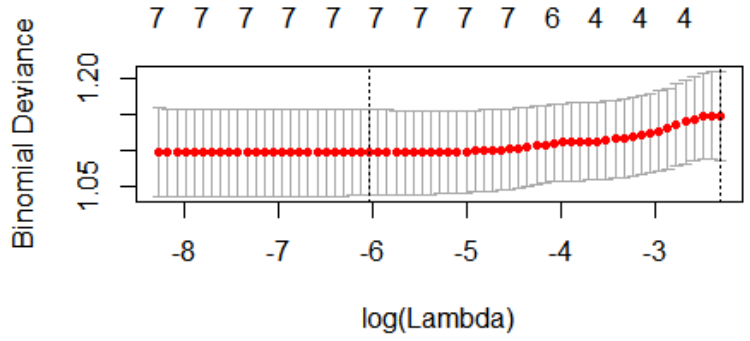
Lasso



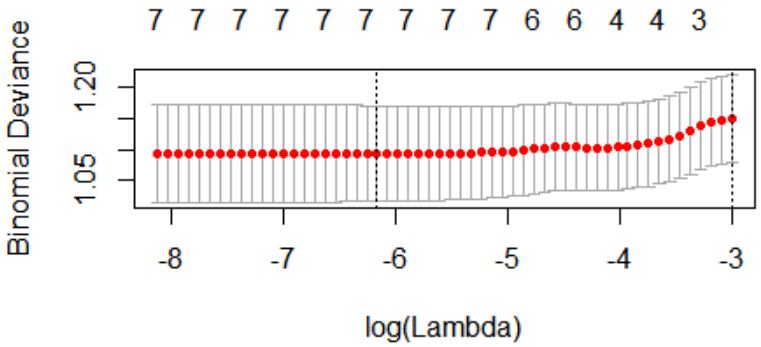
Elastic net

Figure 4.1. Plots of coefficients vs log lambda

In Figure 4.1 the plots show nonzero coefficient estimates as a function of the tuning parameter λ. Lasso and elastic net have the similar plots. In two plots, the values of some coefficients are zero. Thus, the relative variables must be removed due to insignificance in the model.



Elastic net



Lasso

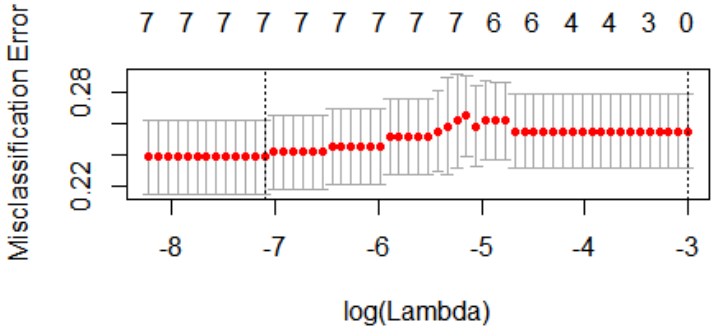
Figure 4.2. Plots of lambda of elastic net vs lasso

In Figure 4.2, the plots show the binomial deviance as a function of the tuning parameter λ in lasso and elastic net regression. The best λ for the model in elastic net and lasso regression are 0.002399396 and 0.002096507.

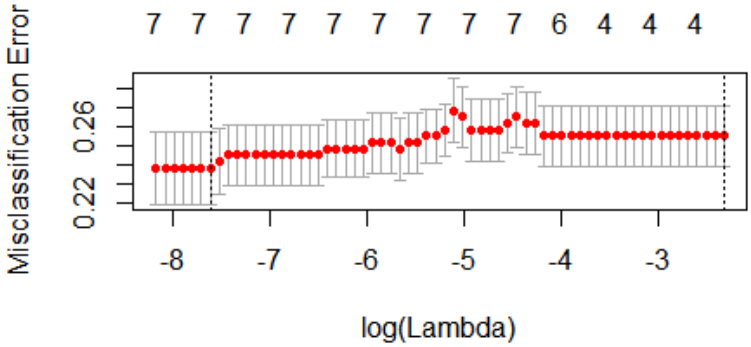
Table 4.3. Coefficient estimates of all predictor variables in elastic net and lasso

Model	***sex***	***trt***	***age***	***baseline2***	***baseline3***	***baseline4***	***baseline5***
elastic net	0.1249	-0.3063	-0.0018	0.9154	1.5977	1.9051	4.6999
Lasso	0.1132	-0.2977	-0.0014	0.9053	1.5925	1.8991	4.7682

From Table 4.3, we could see the values of the coefficients in elastic net and lasso model when the best tuning parameters have been selected by ten-fold cross-validation. In elastic net and lasso, none of coefficient estimate of a predictor variable is equal to zero [6]. The absolute values of coefficients of “baseline5” are the greatest in two models. That means “baseline5” is the most significant variable in two models. Overall, the absolute value of “baseline” is the greatest that demonstrates “baseline” is the most significant value in two models. In contrast, the values of coefficients of “sex”, “age” and “trt” are less than 1 in two models. It indicates these three variables do not affect the response variable “y” significantly.



Lasso



Elastic net

Figure 4.3. Plots of misclassification error vs lambda

From Figure 4.3, we could know that elastic net model has a better fit for the RA dataset compare to lasso model since its values of overall misclassification error are less than lasso.

In conclusion, there is no obvious difference on variable selection of two shrinkage methods. The elastic net fits the RA data set better than lasso due to smaller values of overall misclassification error.

CHAPTER 5: CONCLUSION

The results in last chapter show that the different significant variable selections by applying different models to the RA dataset. The results of all models show that “baseline” is the most significant predictor variable. The variable “trt” has the impact on “y-final RA score” in the logistic model when “y” is defined as the ordinal outcome. Typically, age and gender are considered as risk factors in rheumatoid arthritis research. However, these two factors do not influence the final RA score in this study.

This study illustrates how logistic regression and shrinkage method applied in a small categorical medical dataset. Logistic regression can make use of one or more predictor variables that may be either continuous or categorical. It is usually used for predicting qualitative responses. Lasso and elastic net can make an easy variable selection when the coefficient estimate is shrunk to be exactly zero [7]. But these two models do not shrink any coefficient estimate to zero in this study. Elastic net model is better on model fitting comparing with lasso model. Comparing to shrinkage methods, logistic regression has better performance on variable selection based on P-value [8].

.

REFERENCES

[1] Rheumatoid Arthritis [Online]. Available at: http://www.rheumatology.org/I-Am-A/Patient-Caregiver/Diseases-Conditions/Rheumatoid-Arthritis [Accessed: 15 October 2015].

[2] Rheumatoid Arthritis [Online].
Available at: http://www.arthritis.org/about-arthritis/types/rheumatoid-arthritis [Accessed 15 October 2015].

[3] Logistic Regression [Online].
Available at: https://en.wikipedia.org/wiki/Logistic_regression [Accessed 15 October 2015].

[4] Sundberg, R. (2002). Shrinkage regression. *Encyclopedia of environmetrics*.

[5] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (p. 6). New York: springer.

[6] Archer, K. J., Hou, J., Zhou, Q., Ferber, K., Layne, J. G., & Gentry, A. E. (2014). ordinalgmifs: An R package for ordinal regression in high-dimensional data settings. *Cancer informatics*, 13, 187.

[7] Purposeful selection of variables in logistic regression. Available at:
http://www.nlm.nih.gov/medlineplus/ency/article/003642.htm [Accessed 15 October 2015].

[8] Archer, K. J., & Williams, A. A. A. (2012). L 1 penalized continuation ratio models for ordinal response prediction using high-dimensional datasets. *Statistics in medicine*, 31(14), 1464-1474.

[9] Statistics in Medicine, 1994.
Available at: http://onlinelibrary.wiley.com/journal/10.1002/(ISSN)1097-0258/issues-3. [Accessed: 15 September 2015].