Correlation Adjusted Penalization in Regression Analysis

by

Qi Er (Angela) Tan

A thesis submitted to the Faculty of Graduate Studies of
The University of Manitoba
in partial fulfilment of the requirements for the degree of

Doctor of Philosophy

Department of Statistics
The University of Manitoba
Winnipeg

Abstract

My thesis proposed new types of correlation adjusted penalization methods to address the issue of multicollinearity in regression analysis. The main purpose is to achieve simultaneous shrinkage of parameter estimators and variable selection for multiple linear regression and logistic regression when the predictor variables are highly correlated. The motivation is that when there is serious issue of multicollinearity, the variances of parameter estimators are significantly large. The new correlation adjusted penalization methods shrink the parameter estimators and their variances to alleviate the problem of multicollinearity.

Multicollinearity is an important issue in regression analysis. When the predictor variables in a regression model are highly correlated, their overlapping contributions to the response variable lead to undesirable results when making statistical inference of the response. On one hand, the traditional methods of automatic variable selection often do not work well to produce a satisfactory model. On the other hand, introducing interaction terms into the model makes the model more complex and difficult to apply.

The latest important trend is to apply penalization methods for simultaneous shrinkage and variable selection. In the literature, the following penalization methods are popular: ridge, bridge, LASSO, SCAD, and OSCAR. Few papers have used correlation based penalization methods, and these correlation based methods in the literature do not work when some correlations are either 1 or -1. This means that these correlation based methods fail if at least two predictor variables are perfectly correlated.

We proposed two new types of correlation adjusted penalization methods that work whether or not the predictor variables are perfectly correlated. The types of correlation adjusted penalization methods proposed in the thesis are intuitive and innovative. We investigated important theoretical properties of these new types of penalization methods, including bias, mean squared error, data argumentation and asymptotic properties, and plan to apply them to real data sets in the near future.

Acknowledgements

I would like to thank Dr. Xikui Wang, my supervisor, for his patient guidance and constant support during this research. Without his helpful comments and motivation, this thesis would not have been possible. I thank Dr. Eric Bohm who shared with me his knowledge and ideas, and provided me with the project of hip and knee replacement appropriateness analysis which motivated this research. I also thank Ms. Lynda Loucks for valuable discussions. Moreover, I thank Dr. Smiley Cheng, Dr. Saumen Mandal and Dr. Deli Li for reading my thesis and for giving me constructive criticisms.

I gratefully acknowledge financial supports from the Hip and Knee Institute and the Mathematics of Information Technology and Complex Systems (MITACS). The funding was crucial for me to successfully complete my thesis. I appreciate financial supports from the Faculty of Science and from Dr. Xikui Wang's research grants.

Finally, I thank my family and friends, and am especially grateful to my son, Calvin, for his love and support.

To My Parents

Contents

Chapter 1

Introduction

1.1 Background and motivation

In linear regression analysis, when a large number of predictor variables are introduced to reduce possible modeling biases or there is serious concern of multicollinearity among the predictor variables, variable selection is an important issue.

Conventional methods to select variables include subset selection procedures and stepwise procedures. With a subset selection procedure, several alternative subsets of variables are proposed and compared with each other by means of R^2 criterion, Mallow's C_p criterion, PRESS (Prediction Sum of Squares) criterion, or other criteria.

Automatic variable selection procedures are more common and are offered with most statistical software such as SAS. These include forward selection, backward elimination and stepwise selection. However, there are major drawbacks of automatic selection procedures. For example, the procedures heavily depend on choices of inclusion and exclusion probabilities, and the backward and forward procedures may end up with different best subsets.

On the other hand, suppose multicollinearity is detected and the predictor variables that cause multicollinearity are identified. As discussed by Ryan (2009), multicollinearity may not be a problem if the goal is to use the linear regression model for prediction. However multicollinearity is a problem if we use the linear regression model for description or control.

Multicollinearity implies that predictor variables form some groups. Within each group, predictor variables are highly correlated. One solution to multicollinearity is to remove one or more of the predictor variables within the same group, but deciding which ones to eliminate is a difficult technical issue. A major consequence of multicollinearity is that the parameter estimators and their variances tend to be large. Therefore the inference of the response is highly variable.

The latest and most popular method to address multicollinearity is the use of penalized regression. Essentially the idea is to put constraints on the parameter estimators when estimating the parameters. This consequently put constraints on the variances of the estimators.

The majority of penalization methods put constraints on the parameter estimators only. For example, this includes the methods of ridge, bridge, LASSO, elastic net, SCAD and OSCAR. Few papers have also incorporated empirical correlations in the penalty functions. Intuitively, including empirical correlations should improve the effects of penalization.

Given that all variables are centered and standardized, we characterize the relationship between two predictor variables by a linear equation. Let γ_{ij} be the sample correlation between two predictor variables X_i and X_j. If γ_{ij} is high, then there is multicollinearity involving X_i and X_j. Furthermore, by means of simple linear regression, $\hat{x}_j = \gamma_{ij} x_i$ is the predicted value of X_j based on $X_i = x_i$. If we replace X_j by $\hat{x}_j$ in the multiple linear regression of Y on $X_1, \cdots, X_n$, then we have the terms $\hat{\beta}_i x_i + \gamma_{ij} \hat{\beta}_j x_i$. Our motivation is that the difference $\beta_i - \gamma_{ij}\beta_j$ should be small.

1.2 Objectives and scope of research

The objective of this thesis is to introduce new methods of penalized least squares for multiple linear regression and penalized likelihood estimation for logistic regression that attempts both regression shrinkage and variable selection.

We call these new regularization methods as CAR (Correlation Adjusted Regression) and CAEN (Correlation Adjusted Elastic Net). My motivation is that including empirical correlations in the penalty function may help improving the shrinkage. Moreover including correlations helps to achieve the group effect. By this, we mean that if one predictor variable of a group of highly correlated predictor variables is not statistically significant, then the other predictor variables in the same group tend to be statistically insignificant as well. Hence we wish to leave the whole group out of the model. However, if any predictor variable of a group of highly correlated predictor variables is statistically significant, then the other predictor variables in the same group tend to be statistically significant, so all predictor variables in this group should be kept in the model. Because correlations implicitly connect the predictor variables, including correlations help to either select the whole group of highly correlated predictor variables or to exclude the whole group.

The new types of penalization, CAR and CAEN, are applied to the ordinary least squares regression, logistic regression and LAD (Least Absolute Deviation) regression. We formulate the objectives and derive the estimators, and investigate the properties of the penalized estimators, including bias, mean squared error, data argumentation, and asymptotic properties.

1.3 Structure of the thesis

General backgrounds of regression analysis, including both the ordinary least squares regression and logistic regression, are reviewed in Chapter 2.

In Chapter 3, we review different types of penalization in the literature, including ridge, bridge, LASSO and its extensions, elastic net and correlation based penalties.

We introduce the first new type of penalized regression CAR and investigate its theoretical properties in Chapter 4.

The second new type of penalized regression CAEN is introduced and its theoretical properties are investigated in Chapter 5.

Chapter 6 concludes the thesis with a summary of the achievements and discussion of future research questions and projects.

Chapter 2

Regression analysis

2.1 Introduction

The use of regression analysis has significant applications in medical research and countless other areas, and is an important component of modern data analysis. The central objective is to understand the relationship between a response (or dependent) variable and a set of predictor variables (also known as explanatory variables, regressors, covariates, or independent variables) and to apply the relationship for the purpose of estimating and/or predicting future responses.

There are many important theoretical, practical and computational issues related to regression modeling and inference, including specification of the link function that relates the response variable and predictor variables, estimation of regression parameters in the link function, measure of model performance, diagnostic statistics to assess the modeling assumptions and goodness-of-fit, and remedial methods in the cases of violation of assumptions.

The response variable can be continuous or categorical. Although some philosophical ideas may be similar for different types of response variables, methodologies are different, in particular on the choice of the link function and assessment of goodness-of-fit of the model.

Effective model building is a significant issue. Essentially, we search for the best fitting and most parsimonious model that is practically meaningful and reasonable to describe the relationship between the response and the set of predictor variables. The fit of the model to the data set is determined by measures of goodness-of-fit, and being most parsimonious requires effective methods of model selection.

Multicollinearity is another important issue in multiple regression. Collinearity means a linear relationship exists between two or more predictor variables, while multicollinearity refers to a situation in which two or more predictor variables are highly linearly correlated. The most extreme case is perfect collinearity (or multicollinearity) where the linear correlation between two predictor variables is either 1 or -1. This happens, for example, when two predictor variables X_1 and X_2 satisfy $X_2 = a + bX_1$ for two real numbers a and b.

In the presence of perfect multicollinearity, parameter estimates of the population multiple linear regression model are not unique. In practice, perfect collinearity occurs rarely. However quite often we face the issue of multicollinearity when there are strong linear relationships among two or more predictor variables. This happens when two or more predictor variables contribute more or less to a same characteristic of the subjects. For a matrix A, let A^T be its transpose and A^{-1} be its inverse matrix if it exists. When predictor variables are highly linearly correlated, the most significant consequence is that entries of $(X^TX)^{-1}$ are large, so the predictor variables contribute overlapping and redundant information. Other consequences of multicollinearity are that some predictor variables are not statistically significant but the model is overall significant, and that the usual interpretation of coefficient estimates fails in the presence of multicollinearity and there is high variability of parameter estimators because the estimated variance-covariance matrix has large diagonal entries.

Many methods are available to detect multicollinearity. These include checking for significant change in the parameter estimate when its corresponding predictor variable is added to or removed from the model, checking for insignificance of individual estimators while the model is overall significant, calculating the Variance Inflation Factor (VIF) and carrying out formal multicollinearity tests.

There are several remedies for dealing with multicollinearity. One method is to select a collection of predictor variables that are minimally correlated with each other. This avoids overfitting the regression model and can be normally done with statistical software. However information from other predictor variables is often wasted. Furthermore, there is no clear way of selecting a collection of predictor variables that forms the best subset.

Since omitting predictor variables may result in potential loss of information, another method is to include interaction terms into the model to account for high linear correlation among the predictor variables. There are several problems with this approach. One is that the form of interaction is not unique and must be carefully determined. The other is that the model is much more complex and has too many terms which reduce the degrees of freedom of the inference of the response, and hence reduces the power for predicting and estimating the response.

In recent years, alternative methods have been introduced to deal with multicollinearity. In particular, the method of penalization has becoming popular and useful. This is also known as simultaneous shrinkage and variable selection. We review this area in Chapter 3. General discussion of regression can be found in Ryan (2009) and Kutner et al. (2005).

2.2 Linear statistical models

The ordinary multiple linear regression model is frequently used and has parameters that are easily interpreted. The response variable is continuous and denoted by Y. Let $X_1, X_2, \cdots, X_p$ denote the predictor variables, where p is the number of predictor variables. The relationship between Y and $X_1, X_2, \cdots, X_p$ can be formulated as a linear regression model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon. \tag{2.1}$$

or for each of the n sub-populations, as

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} + \epsilon_i, \quad i = 1, 2, \cdots, n. \tag{2.2}$$

where

- Y_i is the response for the i^{th} of the n sub-populations, $i = 1, 2, \cdots, n$;
- $X_{i1}, X_{i2}, \cdots, X_{ip}$ are the p predictor variables for observation i, $i = 1, 2, \cdots, n$, that determine the i^{th} sub-population.
- $\beta_0, \beta_1, \beta_2, \cdots, \beta_p$ are $p+1$ unknown parameters to be estimated from the data;
- ϵ_i is the random error for the i^{th} sub-population specified by $(X_{i1}, \cdots, X_{ip})$.

In dealing with the regression equation, estimating the parameters, and drawing inference of the responses, the statistical assumptions of the linear regression model include LINE:

- Linearity: $E(Y_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip}$ is a linear function in the parameters $\beta_0, \beta_1, \beta_2, \cdots, \beta_p$;
- Independence: ϵ_i and ϵ_j are independent for $i \neq j$, so Y_i and Y_j are also independent;
- Normality: the random errors $\epsilon_1, \cdots, \epsilon_n$ are normally distributed so that $\epsilon_i \sim N(0, \sigma_i^2)$, implying that Y_i, $i = 1, \cdots, n$, are also normally distributed;

- Equal Variance: the variance of the random error is the same for all subpopulations, $Var(\epsilon_i) = {\sigma_i}^2 = \sigma^2$, for all $i = 1, \cdots, n$, implying that Y_i, $i = 1, \cdots, n$, also have the same variance if we assume that $X_1, \cdots, X_p$ are pre-specified.

Under the assumptions LINE and the values of $x_{i1}, x_{i2}, \cdots, x_{ip}$ of the predictor variables, the random response Y_i follows a normal distribution with mean $\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}$ and variance σ^2.

The linear statistical model can be rewritten in matrix form as

$$Y_i = \begin{pmatrix} 1 & X_{i1} & X_{i2} & \cdots & X_{ip} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix} + \epsilon_i = X_{(i)}\beta + \epsilon_i, \quad i = 1, 2, \cdots, n,$$

where $X_{(i)} = \begin{pmatrix} 1 & X_{i1} & X_{i2} & \cdots & X_{ip} \end{pmatrix}$ and $\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}$, or collectively

$$\begin{aligned} \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} &= \begin{pmatrix} 1 & X_{11} & \cdots & X_{1p} \\ 1 & X_{21} & \cdots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & \cdots & X_{np} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}, \qquad (2.3) \\ or, \qquad Y &= X\beta + \epsilon, \end{aligned}$$

where $\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}$, $\epsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$, and $\mathbf{X} = \begin{pmatrix} 1 & X_{11} & \cdots & X_{1p} \\ 1 & X_{21} & \cdots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & \cdots & X_{np} \end{pmatrix}$ is called the design matrix of the linear model.

The assumptions of Linearity, Independence, Normality, and Equal variance can be collectively expressed as

$$\epsilon \sim N(0, \sigma^2 I_n)$$

where $N(0, \sigma^2 I_n)$ denotes the multivariate normal distribution with mean vector zero and variance-covariance matrix

$$\sigma^2 I_n = \begin{pmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{pmatrix}.$$

Equivalently, the linear model in a matrix form can be rewritten as

$$\mathbf{Y} \sim N(\mathbf{X}\beta, \sigma^2 I_n) \tag{2.4}$$

The first objective of any regression analysis is to find the best fit of the regression model to an observed data set. There are some technical issues. Firstly, the corresponding sample equation that describes the sample data set is of the form

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p.$$

Secondly, there must be a criterion by which we can define the best fit. There are two commonly used criteria: the principle of least squares and the principle of maximum likelihood.

The principle of least squares states that the best fit of the linear model is given by that $\sum_{i=1}^{n} e_i = \sum_{i=1}^{n} (y_i - \hat{y}_i) = 0$ and $\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$ is minimized, where $e_i = y_i - \hat{y}_i$ is called the residual.

The principle of maximum likelihood states that the best fit of the linear model is given by the estimators $\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_p$ that attain the maximum of the likelihood function $\ell(\beta_0, \beta_1, \cdots, \beta_p)$.

2.2.1 Standardized regression

Assume p predictors $X_1^*, X_2^*, \cdots, X_p^*$ and a response Y^* for a linear regression model

$$E(Y^*) = \beta_0^* + \beta_1^* X_1^* + \beta_2^* X_2^* + \cdots + \beta_p^* X_p^*.$$

The observed dataset and summary statistics are displayed as follows:

	Subject				*Sample*	*Sample*
Variable	1	2	$\cdots$	n	*mean*	*standard deviation*
Y^*	y_1^*	y_2^*	$\cdots$	y_n^*	$\overline{y^*}$	S_{y^*}
X_1^*	x_{11}^*	x_{21}^*	$\cdots$	x_{n1}^*	$\overline{x_1^*}$	$S_{x_1^*}$
X_2^*	x_{12}^*	x_{22}^*	$\cdots$	x_{n2}^*	$\overline{x_2^*}$	$S_{x_2^*}$
$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$	$\vdots$
X_p^*	x_{1p}^*	x_{2p}^*	$\cdots$	x_{np}^*	$\overline{x_p^*}$	$S_{x_p^*}$

This means that the original dataset is given by

$$\left\{(y_i^*, x_{i1}^*, \cdots, x_{ip}^*)^T, i = 1, 2, \cdots, n\right\},$$

where T stands for transpose.

We now transform all observations by standardizing. The correlation transformed observations are,

$$y_i = \frac{1}{\sqrt{n-1}} \frac{y_i^* - \overline{y^*}}{S_{y^*}}, \qquad i = 1, 2, \cdots, n,$$

and

$$x_{ik} = \frac{1}{\sqrt{n-1}} \frac{x_{ik}^* - \overline{x_k^*}}{S_{x_k^*}}, \qquad i = 1, 2, \cdots, n, \qquad k = 1, 2, \cdots, p.$$

Then the standardized regression becomes

$$\hat{y} = \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p.$$

There are two major reasons for standardizing. Firstly, the round-off error for least squares estimation (as well as for MLE) is significantly reduced. Secondly, the interpretation of estimated parameters is compatible because the new variables $X_1, X_2, \cdots, X_p$ have no units.

Throughout the rest of the thesis for multiple linear regression with a continuous response, we assume that all variables are standardized.

2.2.2 Method of least squares

With standardized response variable Y and predictor variables $X_1, X_2, \cdots, X_p$, the least squares equation is given as

$$\hat{y} = \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p,$$

where the ordinary least squares (OLS) estimators $\hat{\beta}_1, \cdots, \hat{\beta}_p$ are derived by minimizing

$$OLS = \sum_{i=1}^{n} (y_i - x_i \beta)^2 = (\mathbf{Y} - \mathbf{X}\beta)^T (Y - X\beta),$$

where $X_i = (x_{i1}, x_{i2}, \cdots, x_{ip})$, $\beta = (\beta_1, \cdots, \beta_p)^T$ and

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}, \quad \mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}.$$

In matrix form, the OLS estimator $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2, \cdots, \hat{\beta}_p)^T$ is derived by setting

$$\frac{\partial(OLS)}{\partial\beta} = \frac{\partial\{(\mathbf{Y} - \mathbf{X}\beta)^T(\mathbf{Y} - \mathbf{X}\beta)\}}{\partial\beta} = 0$$

and obtaining

$$\hat{\beta}(OLS) = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}.$$

Furthermore, the variance-covariance matrix of $\hat{\beta}(OLS)$ is given by

$$\begin{aligned} Var(\hat{\beta}) &= Var[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y})] \\ &= \{(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\}Var(\mathbf{Y})\{(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)]^T\} \\ &= \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}. \end{aligned}$$

where σ^2 is unknown and estimated by $MSE = \frac{SSE}{n-p} = \frac{e^T e}{n-p}$, and $e = y - \hat{y} = (y_1 - \hat{y}_1, \cdots, y_n - \hat{y}_n)^T$ is the vector of residuals.

The vector of predicted values of the response is given by

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\beta} = [\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T]\mathbf{Y} = H\mathbf{Y},$$

where $H = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ is called the hat matrix. Clearly H is idempotent (i.e., $H = H^2$) and symmetric (i.e., $H^T = H$).

2.2.3 Method of maximum likelihood

Maximum likelihood estimation is a general technique for estimating parameters and drawing statistical inferences in a variety of situations. Based on the available data, we wish to estimate the parameters $\beta_1, \cdots, \beta_p$ that make the probability of observing the data as high as possible. This is called the principle of Maximum Likelihood Estimation (MLE).

For fixed values of $\mathbf{X}$ that fall within the range of the data, the probability model for the response $\mathbf{Y}$ is given by $\mathbf{Y} \sim N(\mathbf{X}\beta, \sigma^2 I_n)$, and the probability density of a normal distribution is

$$L(\beta^T, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} exp\left[-\frac{(\mathbf{Y}-\mathbf{X}\beta)^T(\mathbf{Y}-\mathbf{X}\beta)}{2\sigma^2}\right]$$

which is called the likelihood function of (β^T, σ^2), where $exp(u) = e^u$ is the natural exponential function. The log-likelihood function, i.e. the natural logarithm of the likelihood function, is

$$\begin{aligned} \ell(\beta^T, \sigma^2) &= logL(\beta^T, \sigma^2) \\ &= -\frac{n}{2}log(2\pi) - \frac{n}{2}log(\sigma^2) - \frac{1}{2\sigma^2}(\mathbf{Y}-\mathbf{X}\beta)^T(\mathbf{Y}-\mathbf{X}\beta). \end{aligned} \tag{2.5}$$

Taking partial derivatives with respect to the parameters and setting them equal to zero, we get the maximum likelihood estimators $\hat{\beta}_1, \cdots, \hat{\beta}_p$. We note that with respect to β, maximizing ℓ is equivalent to minimize the Ordinary Least Squares (OLS). The partial derivatives are

$$\frac{\partial \ell(\beta^T, \sigma^2)}{\partial \beta^T} = -\frac{1}{2\sigma^2}(2\mathbf{X}^T\mathbf{X}\beta - 2\mathbf{X}^T\mathbf{Y}),$$

$$\frac{\partial \ell(\beta^T, \sigma^2)}{\partial \sigma^2} = -\frac{n}{2}\left(\frac{1}{\sigma^2}\right) + \frac{1}{\sigma^4}(\mathbf{Y} - \mathbf{X}\beta)^T(\mathbf{Y} - \mathbf{X}\beta).$$

Setting these partial derivatives to zero and solving the equations, the maximum likelihood estimator $\hat{\beta}$ satisfies

$$(\mathbf{X}^T\mathbf{X})\hat{\beta} = \mathbf{X}^T\mathbf{Y},$$

where $\hat{\beta} = \begin{pmatrix} \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_p \end{pmatrix}$. Provided that $|\mathbf{X}^T\mathbf{X}| \neq 0$, the solution is

$$\hat{\beta}_n(MLE) = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}. \tag{2.6}$$

$$\hat{\sigma^2} = \frac{(\mathbf{Y} - \mathbf{X}\hat{\beta})^T(\mathbf{Y} - \mathbf{X}\hat{\beta})}{n} = \frac{e^T e}{n}, \tag{2.7}$$

where $e = \mathbf{Y} - \mathbf{X}\hat{\beta}$ is the residual vector. Clearly the MLE $\hat{\beta}_n(MLE) = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$ is the same as the principle of least squares estimators of $\beta_1, \cdots, \beta_p$.

The well known Gauss-Markov Theorem states that the MLE $\hat{\beta}$ is BLUE, the Best, Linear, Unbiased Estimator. First of all, $\hat{\beta}$ is unbiased in that $E(\hat{\beta}) = \beta$. Secondly, $\hat{\beta}$ is a linear estimator because $\hat{\beta} = ((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)\mathbf{Y}$ is a linear function of Y. Lastly, $\hat{\beta}$ is the best among the class of all linear unbiased estimators of β in the sense that the variances of $\hat{\beta}_i$, $i = 1, 2, \cdots, p$, are the smallest.

2.3 Logistic regression

When the response is binary (i.e., dichotomous), it is no longer reasonable to model the conditional expectation of the binary response as a linear function of the parameter. If this were the case, then we would have $E(Y|x_1, \cdots, x_p) = \beta_1 x_1 + \cdots + \beta_p x_p$. Clearly, $E(Y|x_1, \cdots, x_p)$ takes only values between 0 and 1, but the right hand side can take any real value. So we need to transform $\beta_1 x_1 + \cdots + \beta_p x_p$ to a value between 0 and 1. The most popular function is the logistic function $\frac{e^x}{1+e^x}$, and hence the name logistic regression. To be specific, the logistic regression is given by the model

$$E(Y|x_1, \cdots, x_p) = \frac{e^{\beta_0+\beta_1 x_1+\cdots+\beta_p x_p}}{1+e^{\beta_0+\beta_1 x_1+\cdots+\beta_p x_p}} \tag{2.8}$$

or

$$log \frac{\pi(x_1, \cdots, x_p)}{1-\pi(x_1, \cdots, x_p)} = \beta_0 + \beta_1 x_1 + \cdots, \beta_p x_p \tag{2.9}$$

where $\pi(x_1, \cdots, x_p) = E(Y|x_1, \cdots, x_p)$ and $g(z) = log\left(\frac{z}{1-z}\right)$ is called the logit transformation.

Recall that for multiple linear regression, we have $Y_i = E(Y_i|x_{i1}, \cdots, x_{ip}) + \epsilon_i$ or simply $Y = E(Y|x_1, \cdots, x_p) + \epsilon$, where ϵ is random and normally distributed with mean 0 and common variance σ^2.

However for logistic regression, if we have

$$Y = E(Y|x_1, \cdots, x_p) + \epsilon = \pi(x_1, \cdots, x_p) + \epsilon.$$

then ϵ takes value $1 - \pi(x_1, \cdots, x_p)$ with probability $\pi(x_1, \cdots, x_p)$ and value $-\pi(x_1, \cdots, x_p)$ with probability $1 - \pi(x_1, \cdots, x_p)$.

Let's write $x_i = (x_{i1}, x_{i2}, \cdots, x_{ip})$ and $\pi(x_{i1}, x_{i2}, \cdots, x_{ip}) = \pi(x_i)$. Then the likelihood function is

$$L(\beta_0, \beta_1, \cdots, \beta_p) = \prod_{i=1}^{n} \left\{ [\pi(x_i)]^{y_i} [1 - \pi(x_i)]^{1-y_i} \right\}, \tag{2.10}$$

and the log-likelihood function is

$$\ell(\beta_0, \beta_1, \cdots, \beta_p) = \sum_{i=1}^{n} \{ y_i log[\pi(x_i)] + (1 - y_i) log[1 - \pi(x_i)] \}. \tag{2.11}$$

Setting the partial derivatives with respect to $\beta_0, \beta_1, \cdots, \beta_p$ to 0, we have the likelihood equations

$$\begin{aligned}
\frac{\partial \ell(\beta)}{\partial \beta_0} &= \sum_{i=1}^{n} e_i = \sum_{i=1}^{n} [y_i - \pi(x_i)] = 0, \\
\frac{\partial \ell(\beta)}{\partial \beta_1} &= \sum_{i=1}^{n} x_{i1} e_i = \sum_{i=1}^{n} x_{i1} [y_i - \pi(x_i)] = 0, \\
&\vdots \\
\frac{\partial \ell(\beta)}{\partial \beta_p} &= \sum_{i=1}^{n} x_{ip} e_i = \sum_{i=1}^{n} x_{ip} [y_i - \pi(x_i)] = 0.
\end{aligned}$$

Note that these are similar to the likelihood equations for multiple linear models, however, these likelihood equations are much more difficult to solve than the ones for multiple linear regression. Thankfully, most statistical software could produce estimates of logistic regression and calculate $\hat{\pi}(x_i) = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p}}$.

Chapter 3

Penalized regression

3.1 Introduction

When there is multicollinearity among the predictor variables $X_1, X_2, \cdots, X_p$, the determinant of the matrix $\mathbf{X}^T\mathbf{X}$ is small where

$$\mathbf{X} = \begin{pmatrix} X_{11} & X_{12} & \cdots & X_{1p} \\ X_{21} & X_{22} & \cdots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & \cdots & X_{np} \end{pmatrix}$$

is the design matrix. Consequently the entries of the inverse matrix $(\mathbf{X}^T\mathbf{X})^{-1}$ are fairly large. The least squares estimators (also the maximum likelihood estimators) of parameters, given by $\hat{\beta}_n(OLS) = (\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{Y})$, are large. More significantly, since the variance-covariance matrix of $\hat{\beta}_n(OLS)$ is $\sigma^2(\hat{\beta}_n(OLS)) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$, the estimated parameters $\hat{\beta}_n(OLS)$ are subject to large variability. Hence, the prediction or estimation of the response, which is a linear function of the estimated parameters, will have large variability. This will adversely affect the quality of estimation or prediction of the response.

Regression is one of the most useful statistical methods for data analysis. However, there are many practical problems and computational issues, such as multicollinearity and high dimensionality, that challenge the regression analysis. To deal with these challenges, variable selection and shrinkage estimation are becoming important and popular. The traditional approach of automatic selection (such as forward selection, backward elimination and stepwise selection) and best subset selection are often computationally expensive and may not necessarily produce the best model.

The method of penalized least squares (PLS), which is equivalent to penalized maximum likelihood, helps to deal with the issue of multicollinearity by putting constraints on the values of the estimated parameters. A wonderful consequence is that the entries of the variance-covariance matric is also significantly reduced.

In general, the PLS is to minimize $OLS = (\mathbf{Y} - \mathbf{X}\beta)^T(\mathbf{Y} - \mathbf{X}\beta)$ subject to $Pen(\beta) \leq t$, where $Pen(\beta)$ is a specific penalty function of $\beta = (\beta_1, \cdots, \beta_p)^T$ and t is a tuning parameter. This constrained optimization problem is equivalent to the Lagrangian optimization which minimizes

$$PLS = OLS + Penalty = (\mathbf{Y} - \mathbf{X}\beta)^T(\mathbf{Y} - \mathbf{X}\beta) + \lambda Pen(\beta),$$

where λ is a tuning parameter and controls the strength of shrinkage. For example, when $\lambda = 0$, no penalty is applied and we have the ordinary least squares regression. When λ gets larger, more weight is given to the penalty term. Desirable properties of penalization include variable selection and grouping effect.

That is, by penalization it is hoped that the variables that are truly statistically significant are selected into the model, and highly correlated predictor variables should be selected all together or excluded all together.

Suppose that the data set consists of n observations: $\{(y_i, x_i)^T, i = 1, 2, \cdots, n\}$, where T stands for transpose, y_i is the response and $x_i = (x_{i,1}, x_{i,2}, \cdots, x_{i,p})$ is the vector of predictors for the i^{th} subject. We assume that all data are standardized so that $\sum_{i=1}^n y_i = 0, \sum_{i=1}^n x_{i,j} = 0$ and $\sum_{i=1}^n x_{i,j}^2 = 1$ for $j = 1, 2, \cdots, p$. Both the multiple linear regression $\mathbf{Y} = \mathbf{X}\beta + \epsilon$ and the logistic regression $P(Y_i = 1) = \frac{1}{1+e^{-X_i\beta}}$ are considered, where $\mathbf{X} = (X_1, X_2, \cdots, X_p)$ is the design matrix. The standard assumptions for the linear models are LINE: Linearity of the model, Independence of $\epsilon_i, i = 1, 2, \cdots, n$, Normality of $\epsilon_i \sim N(0, \sigma_i^2)$ and Equal variance $\sigma_i^2 = \sigma^2$ for all $i = 1, 2, \cdots, n$.

Many different forms of the penalty functions have been introduced in the literature, including ridge penalty, bridge penalty, LASSO (Least Absolute Shrinkage and Selection Operator) and its generalizations, elastic net, SCAD (Smoothly Clipped Absolute Deviation), and correlation based penalties.

3.2 Ridge regression

The ridge regression, introduced by Hoerl and Kennard (1970), may be the earliest idea of using penalized least squares. The objective is to minimize

$$RidgeLS = OLS + \lambda \sum_{j=1}^{p} \beta_j^2 = (\mathbf{Y} - \mathbf{X}\beta)^T(\mathbf{Y} - \mathbf{X}\beta) + \lambda\beta^T\beta.$$

This is also referred as L_2 penalized least squares.

Instead of having $\hat{\beta}_n(OLS) = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$ for the ordinary least squares regression, ridge regression estimators are derived as

$$\hat{\beta}_n(Ridge) = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{Y}$$

where $\mathbf{I}$ is the identity matrix of size $p \times p$.

When $\lambda = 0$, this becomes the ordinary least squares estimation. When $\lambda > 0$, ridge regression estimators are biased but tend to be less variable, therefore have smaller mean squared errors for appropriately chosen values of λ.

Ridge regression estimators are robust and could provide good estimates of the mean responses or individual responses. However, a major limitation is that the usual inference procedures are not applicable and exact distributional properties are unknown.

One straightforward extension is the generalized ridge estimator

$$\hat{\beta}_n(GenRidge) = (\mathbf{X}^T\mathbf{X} + \mathbf{K})^{-1}\mathbf{X}^T\mathbf{Y}$$

where K is a diagonal matrix with possibly different diagonal elements $k_i \geq 0, i = 1, 2, \cdots, n$. See for example Alheety and Ramanathan (2009). Ridge penalty is also applied to logistic regression, see Le Cessie and Van Houwelingen (1992), Barker and Brown (2001), and Mansson and Shukur (2011).

3.3 Bridge regression

Frank and Friedman (1993) extended ridge regression to bridge regression by generalizing the penalty function to

$$Pen(\beta) = \lambda \sum_{j=1}^{p} |\beta_j|^{\gamma},$$

where $\gamma \geq 0$ is also a tuning parameter. The objective is to minimize

$$BridgeLS = (\mathbf{Y} - \mathbf{X}\beta)^T(\mathbf{Y} - \mathbf{X}\beta) + \lambda \sum_{j=1}^{p} |\beta_j|^{\gamma}.$$

Bridge regression is also called L_γ penalized regression.

3.4 $L_{1/2}$ regularization

Xu et al. (2010) examined a special bridge regression with the penalty

$$Pen(\beta) = \lambda \sum_{j=1}^{p} |\beta_j|^{1/2}.$$

The objective is to minimize

$$HalfLS = \frac{1}{n}(Y - X\beta)^T(Y - X\beta) + \lambda \sum_{j=1}^{p} |\beta_j|^{1/2}.$$

This is also named $L_{1/2}$ penalized regression.

3.5 LASSO

LASSO (Least Absolute Shrinkage and Selection Operator) was introduced by Tibshirani (1996), also known as L_1 penalized regression. The penalty function is

$$Pen(\beta) = \lambda \sum_{j=1}^{p} |\beta_j|.$$

The objective is to minimize

$$\begin{aligned} LASSOLS &= (\mathbf{Y} - \mathbf{X}\beta)^T(\mathbf{Y} - \mathbf{X}\beta) + \lambda \sum_{j=1}^{p} |\beta_j| \\ &= (\mathbf{Y} - \mathbf{X}\beta)^T(\mathbf{Y} - \mathbf{X}\beta) + \lambda \mathbf{V}^T \beta, \end{aligned}$$

where $\mathbf{V}$ is a column vector with i^{th} element being 1 if the sign of the corresponding parameter β_i is positive and -1 if β_i is negative. Although $\mathbf{V}$ depends on the unknown parameters through their signs, in practice based on theory or empirical evidence, we might be able to determine the signs of the unknown parameters in advance and so $\mathbf{V}$ could be regarded as being pre-determined.

The estimators that minimize $LASSOLS$ could be derived as

$$\hat{\beta}_n(LASSO) = (\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{Y} - \frac{\lambda}{2}\mathbf{V}).$$

LASSO generally results in simultaneous shrinkage and variable selection. This means that some estimators become identically zero if their corresponding parameters are zero. This is called the oracle property. However, a major drawback of LASSO is that if there is a group of highly correlated predictor variables in the regression model, LASSO tends to arbitrarily select only one predictor variable from this group.

Although LASSO has demonstrated good performance in many cases, limitations still exist, see for example Tibshirani (1996) and Kyung et al. (2010). In fact, ridge regression dominates the LASSO in prediction performance when there are severe multicollinearity presence among predictor variables. Asymptotic properties of LASSO estimators are discussed in Knight and Fu (2000).

To deal with the limitation of LASSO, extensions and variants of LASSO have been proposed. Tibshirani et al. (2005) introduced the fused LASSO, given by

$$FLASSOPen(\beta) = (\mathbf{Y} - \mathbf{X}\beta)^T(\mathbf{Y} - \mathbf{X}\beta) + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=2}^{p} |\beta_j - \beta_{j-1}|.$$

where λ_1 and λ_2 are tuning parameters. Zou (2006) developed the adaptive LASSO and Park and Casella (2008) proposed the Bayesian LASSO.

3.6 Elastic net regression

Combining the L_1 and L_2 penalties, Zou and Hastie (2005) introduced the elastic net by minimizing

$$\begin{aligned} ENLS &= (\mathbf{Y} - \mathbf{X}\beta)^T(\mathbf{Y} - \mathbf{X}\beta) + \lambda_1 \sum_{i=j}^{p} |\beta_j| + \lambda_2 \sum_{i=j}^{p} \beta_j^2 \\ &= (\mathbf{Y} - \mathbf{X}\beta)^T(\mathbf{Y} - \mathbf{X}\beta) + \lambda_1 \mathbf{V}^T\beta + \lambda_2 \beta^T\beta \end{aligned}$$

with two tuning parameters λ_1 and λ_2. Later Zou and Zhang (2009) proposed the adaptive elastic net. Li and Lin (2010) introduced the Bayesian elastic net. The estimators that minimize $ENLS$ are

$$\hat{\beta}_n(ENLS) = (\mathbf{X}^T\mathbf{X} + \lambda_2\mathbf{I})^{-1}(\mathbf{X}^T\mathbf{Y} - \frac{\lambda_1}{2}\mathbf{V})$$

where $\mathbf{V}$ is introduced in the section of LASSO.

3.7 SCAD regression

Fan and Li (2001) suggested a SCAD (Smoothly Clipped Absolute Deviation) penalty given by

$$Pen(\beta) = \lambda\left[I(\theta \leq \lambda) + \frac{(a\lambda - \theta)_+}{(a-1)\lambda} T(\theta > \lambda)\right]$$

for some $a > 2$ and $\theta > 0$.

Large sample properties of SCAD estimators are studies in Kwon and Kim (2012).

3.8 OSCAR regression

Bondell and Reich (2008) introduced the OSCAR (octagonal shrinkage and clustering algorithm for regression) penalty

$$Pen(\beta) = \lambda\left[\sum_{j=1}^{p} |\beta_j| + c\sum_{j<k} \max\{|\beta_j|, |\beta_k|\}\right]$$

where $c \geq 0$ is a tuning parameter. The objective is to minimize

$$\begin{aligned} OSCARLS &= (Y - X\beta)^T (Y - X\beta) + \lambda \left[\sum_{j=1}^{p} |\beta_j| + c \sum_{j<k} \max\{|\beta_j|, |\beta_k|\} \right] \\ &= (Y - X\beta)^T (Y - X\beta) + \lambda \sum_{j=1}^{n} [c(j-1)+1] |\beta|_{(j)} \end{aligned}$$

where $|\beta|_{(1)} \leq |\beta|_{(2)} \leq \cdots \leq |\beta|_{(p)}$.

3.9 Correlation based penalization

Tutz and Ulbricht (2009) introduced a correlation based penalty, which minimizes

$$CPLS = (\mathbf{Y} - \mathbf{X}\beta)^T (\mathbf{Y} - \mathbf{X}\beta) + \lambda \sum_{j=1}^{p-1} \sum_{k>j} \left[\frac{(\beta_j - \beta_k)^2}{1 - r_{j,k}} + \frac{(\beta_j + \beta_k)^2}{1 + r_{j,k}} \right],$$

where $r_{j,k}$ is the empirical correlation between the predictors X_j and X_k. This penalty is also investigated in Ulbricht and Tutz (2008). Moreover, Anbari and Mkhadri (2008) introduced the elastic corr-net by minimizing

$$\begin{aligned} ECNLS &= (\mathbf{Y} - \mathbf{X}\beta)^T (\mathbf{Y} - \mathbf{X}\beta) + \lambda_1 \sum_{j=1}^{n} |\beta_j| \\ & \quad + \lambda_2 \sum_{j=1}^{p-1} \sum_{k>j} \left[\frac{(\beta_j - \beta_k)^2}{1 - r_{j,k}} + \frac{(\beta_j + \beta_k)^2}{1 + r_{j,k}} \right]. \end{aligned}$$

3.10 Summary

The motivation for using penalized regression is that in the presence of perfect multicollinearity, the ordinary least squares estimates are not unique. However with penalized least squares, the estimates are unique by choosing appropriate tuning parameters. Similarly, without penalization, the ordinary least squares estimators are subject to high variabilities when multicollinearity exists. With penalization, the variances of the estimators are controlled.

A disadvantage of the ridge regression is that the interpretation is not easy since the final model includes all input variables. There are methods to do variable selection and shrinkage estimation simultaneously, such as the LASSO and elastic net methods. The ridge regression only shrinks the estimates to 0, but LASSO also selects variable automatically. One very interesting and important property of LASSO is that the predictive model is sparse (i.e., the estimators are exactly 0 if the corresponding parameters are truly 0). The elastic net improves on both ridge regression and LASSO and includes the groups of highly correlated predictors, while LASSO selects only one of each group.

Fan and Li (2001) suggested three desirable properties for a good penalty function: unbiasedness (i.e., the estimator is nearly unbiased when the true unknown parameter is large), sparsity (i.e., the parameter is estimated to be zero when the true unknown parameter is zero) and continuity (the estimator is continuous in data). Various advantages and disadvantages of each of these penalization methods and many theoretical as well as practical performances are investigated in the literature. In summary, ridge regression performs well when the predictors are highly correlated. Bridge regression includes both ridge and LASSO as special cases and produces sparse models.

The correlation based penalty has the group effect. That is, a group of highly correlated predictors are either all selected together into the model or left out altogether. However the above introduced penalty does not work when the correlation $r_{j,k}$ is either 1 or -1. In the next two chapters, we extend the correlation based penalty to two new forms to deal with this drawback. One motivation for our new penalty is that if X_j and X_k are highly correlated, then the prediction of X_j using X_k is given by $\hat{x}_k = r_{j,k}x_j$. We minimize $\beta_j - r_{j,k}\beta_k$ to equalize the contributions by both X_j and X_k.

Chapter 4

Correlation adjusted regression (CAR)

4.1 Introduction

In this chapter, we propose two new types of regularization method for simultaneous shrinkage and variable selection, we name them Correlation Adjusted Regression (CAR). They could be regarded as a data-adjusted extension of ridge regression. Some theoretical results for multiple linear regression, logistic regression and LAD (Least Absolute Deviation) regression are derived. We would show that the CAR least squares for multiple linear regression is reduced to the OLS after applying argumentation to the data set, and the penalized estimators for CAR logistic regression follow asymptotically a normal distribution. Similar results are also derived for the LAD regression.

4.2 CAR for linear models

We first discuss two types of CAR for multiple linear regression. Recall that we assume centered and standardized observations and the ordinary least squares regression that minimizes

$$OLS = \sum_{i=1}^{n}(y_i - x_i\beta)^2 = (\mathbf{Y} - \mathbf{X}\beta)^T(\mathbf{Y} - \mathbf{X}\beta)$$

where

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}, \quad \mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix},$$

$\beta = (\beta_1, \beta_2, \cdots, \beta_p)^T$ is the transpose of the vector of unknown regression parameters, and $x_i = (x_{i1}, x_{i2}, \cdots, x_{ip})$ is the i^{th} row of the design matrix. Recall also that the least squares estimator of β is $\hat{\beta}_n(OLS) = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$ and its variance-covariance matrix is $Var(\hat{\beta}_n(OLS)) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$ where σ^2 is the population variance of the regression model.

4.2.1 Formulation of the problem

We define two types of CAR by incorporating empirical correlation coefficients in the penalty function. The first type is defined by the correlation adjusted least squares

$$CARLS_1 = (\mathbf{Y} - \mathbf{X}\beta)^T(\mathbf{Y} - \mathbf{X}\beta) + \lambda\left[\sum_{j=1}^{p-1}(\beta_j - r_{j,j+1}\beta_{j+1})^2 + \beta_p^2\right]$$

where $r_{j,k}$ is the sample correlation between the predictor variables X_j and X_k. The objective is to find $\hat{\beta}_n(CAR_1)$ that minimizes $CARLS_1$.

For the first type, define the matrix

$$D_1 = \begin{pmatrix} 1 & -r_{1,2} & 0 & \cdots & 0 & 0 \\ 0 & 1 & -r_{2,3} & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & -r_{p-1,p} \\ 0 & 0 & 0 & \cdots & 0 & 1 \end{pmatrix},$$

then $\sum_{j=1}^{p-1}(\beta_j - r_{j,j+1}\beta_{j+1})^2 + \beta_p^2 = \beta^T D_1^T D_1 \beta = \beta^T W_1 \beta$, where $W_1 = D_1^T D_1$. Clearly W_1 is a real symmetric $p \times p$ matrix and positive semi-definite because $\beta^T W_1 \beta \geq 0$ for any vector β. Therefore W_1 admits a Cholesky's decomposition $W_1 = C_1 C_1^T$, where C_1 is a triangular matrix.

Moreover if $r_{j,j+1} = 0$ for all $j = 1, 2, \cdots, p-1$, then the first type CAR becomes ridge regression and hence ridge regression is a special case of CAR.

The second type is defined by the correlation adjusted least squares

$$CARLS_2 = OLS + \lambda \left[\sum_{j=1}^{p-1} \sum_{k>j} (\beta_j - r_{j,k}\beta_k)^2 + \beta_p^2 \right].$$

The objective is to find $\hat{\beta}_n(CAR_2)$ that minimizes $CARLS_2$.

We similarly define the matrix of the second type

$$D_2 = \begin{pmatrix} 1 & -r_{1,2} & 0 & 0 & \cdots & 0 & 0 \\ 1 & 0 & -r_{1,3} & 0 & \cdots & 0 & 0 \\ 1 & 0 & 0 & -r_{1,4} & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & 0 & \cdots & 0 & -r_{1,p} \\ 0 & 1 & -r_{2,3} & 0 & \cdots & 0 & 0 \\ 0 & 1 & 0 & -r_{2,4} & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 1 & 0 & 0 & \cdots & 0 & -r_{2,p} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 1 & -r_{p-1,p} \\ 0 & 0 & 0 & 0 & \cdots & 0 & 1 \end{pmatrix},$$

then $\sum_{j=1}^{p-1}\sum_{k>j}(\beta_j - r_{j,k}\beta_k)^2 + \beta_p^2 = \beta^T D_2^T D_2 \beta = \beta^T W_2 \beta$, where $W_2 = D_2^T D_2$. Again W_2 is a real symmetric $p \times p$ matrix and positive semi-definite because $\beta^T W_2 \beta \geq 0$ for any vector β. Therefore again W_2 admits a Cholesky's decomposition $W_2 = C_2 C_2^T$, where C_2 is a triangular matrix.

Putting both types together, we minimize $CARLS = OLS + \lambda\beta^T W\beta$ where W can be either W_1 or W_2. The derivative of $CARLS$ with respect to β is given as the column vector

$$\begin{aligned} \frac{d(CARLS)}{d\beta} &= -2[\mathbf{X}^T\mathbf{Y} - (\mathbf{X}^T\mathbf{X})\beta] + 2\lambda W\beta \\ &= -2[\mathbf{X}^T\mathbf{Y} - (\mathbf{X}^T\mathbf{X} + \lambda W)\beta]. \end{aligned}$$

Therefore the penalized estimator of β is

$$\hat{\beta}_n(CAR) = (\mathbf{X}^T\mathbf{X} + \lambda W)^{-1}(\mathbf{X}^T\mathbf{Y}).$$

This gives a linear estimator $\hat{\beta}_n(CAR) = [(\mathbf{X}^T\mathbf{X} + \lambda W)^{-1}\mathbf{X}^T]\mathbf{Y}$ of $\mathbf{Y}$. The predicted value of $\mathbf{Y}$ is given by $\hat{y} = \mathbf{X}\hat{\beta}_n(CAR) = H\mathbf{Y}$ where $H = \mathbf{X}(\mathbf{X}^T\mathbf{X} +$

$\lambda W)^{-1}\mathbf{X}^T$ is a symmetric hat matrix. This result shows that the penalized estimator of the least squares regression with a correlation adjusted penalty exists.

Comparing with $\hat{\beta}_n(Ridge) = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}(\mathbf{X}^T\mathbf{Y})$, we see that CAR estimator can be regarded as an extension of ridge by replacing the identity matrix $\mathbf{I}$ with the matrix W.

4.2.2 Some properties

We derive the bias of the estimator $\hat{\beta}_n(CAR)$. For this,

$$\begin{aligned}
E(\hat{\beta}_n(CAR)) &= (\mathbf{X}^T\mathbf{X} + \lambda W)^{-1}(\mathbf{X}^T\mathbf{X})\beta \\
&= (\mathbf{X}^T\mathbf{X} + \lambda W)^{-1}(\mathbf{X}^T\mathbf{X} + \lambda W - \lambda W)\beta \\
&= \beta - \lambda(\mathbf{X}^T\mathbf{X} + \lambda W)^{-1}W\beta.
\end{aligned}$$

Therefore the bias is

$$Bias = E(\hat{\beta}_n(CAR)) - \beta = -\lambda(\mathbf{X}^T\mathbf{X} + \lambda W)^{-1}W\beta$$

and $\hat{\beta}_n(CAR)$ is a biased estimator of β.

The variance of the estimator is then

$$\begin{aligned}
Var(\hat{\beta}_n(CAR)) &= Var[(\mathbf{X}^T\mathbf{X} + \lambda W)^{-1}\mathbf{X}^T)\mathbf{Y}] \\
&= [(\mathbf{X}^T\mathbf{X} + \lambda W)^{-1}\mathbf{X}^T]Var(\mathbf{Y})[\mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda W)^{-1}] \\
&= (\mathbf{X}^T\mathbf{X} + \lambda W)^{-1}\mathbf{X}^T\sigma^2\mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda W)^{-1}
\end{aligned}$$

and the mean squared error of the estimator is

$$MSE(\hat{\beta}_n(CAR)) = Bias^T Bias + trace[Var(\hat{\beta}_n(CAR))],$$

where *trace* is the trace of a matrix, that is the sum of the main diagonal elements of a square matrix.

The following Theorem shows that when the tuning parameter λ gets large, the variance-covariance matrix of the estimator approaches the 0 matrix.

Theorem 4.2.1. For the estimator $\hat{\beta}_n(CAR)$, $\lim_{\lambda\to\infty} Var(\hat{\beta}_n(CAR)) = 0$, the 0 matrix.

Proof.

$$
\begin{aligned}
& Var(\hat{\beta}_n(CAR)) \\
= \; & (\mathbf{X}^T\mathbf{X} + \lambda W)^{-1}\mathbf{X}^T\sigma^2\mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda W)^{-1} \\
= \; & \frac{\sigma^2}{\lambda}\left(\frac{\mathbf{X}^T\mathbf{X}}{\lambda} + W\right)^{-1}\frac{\mathbf{X}^T\mathbf{X}}{\lambda}\left(\frac{\mathbf{X}^T\mathbf{X}}{\lambda} + W\right)^{-1} \\
= \; & \frac{\sigma^2}{\lambda}\left(\frac{\mathbf{X}^T\mathbf{X}}{\lambda} + W\right)^{-1}\left(\frac{\mathbf{X}^T\mathbf{X}}{\lambda} + W - W\right)\left(\frac{\mathbf{X}^T\mathbf{X}}{\lambda} + W\right)^{-1} \\
= \; & \frac{\sigma^2}{\lambda}\left[\left(\frac{\mathbf{X}^T\mathbf{X}}{\lambda} + W\right)^{-1} - \left(\frac{\mathbf{X}^T\mathbf{X}}{\lambda} + W\right)^{-1}(W^{-1})^{-1}\left(\frac{\mathbf{X}^T\mathbf{X}}{\lambda} + W\right)^{-1}\right] \\
= \; & \frac{\sigma^2}{\lambda}\left[\left(\frac{\mathbf{X}^T\mathbf{X}}{\lambda} + W\right)^{-1} - \left(W^{-1}\frac{\mathbf{X}^T\mathbf{X}}{\lambda} + I\right)^{-1}\left(\frac{\mathbf{X}^T\mathbf{X}}{\lambda} + W\right)^{-1}\right] \\
= \; & \frac{\sigma^2}{\lambda}\left[\left(\frac{\mathbf{X}^T\mathbf{X}}{\lambda} + W\right)^{-1} - \left(\left(\frac{\mathbf{X}^T\mathbf{X}}{\lambda} + W\right)\left(W^{-1}\frac{\mathbf{X}^T\mathbf{X}}{\lambda} + I\right)\right)^{-1}\right] \\
= \; & \frac{\sigma^2}{\lambda}\left[\left(\frac{\mathbf{X}^T\mathbf{X}}{\lambda} + W\right)^{-1} - \left(\frac{\mathbf{X}^T\mathbf{X}}{\lambda}W^{-1}\frac{\mathbf{X}^T\mathbf{X}}{\lambda} + 2\frac{\mathbf{X}^T\mathbf{X}}{\lambda} + W\right)^{-1}\right].
\end{aligned}
$$

Since the elements of $\mathbf{X}$ and $\mathbf{X}^T\mathbf{X}$ are bounded,

$$
\lim_{\lambda\to\infty}\left(\frac{\mathbf{X}^T\mathbf{X}}{\lambda} + W\right) = W
$$

and

$$\lim_{\lambda\to\infty}\left(\frac{\mathbf{X}^T\mathbf{X}}{\lambda}W^{-1}\frac{\mathbf{X}^T\mathbf{X}}{\lambda}+2\frac{\mathbf{X}^T\mathbf{X}}{\lambda}+W\right)=W,$$

and

$$\lim_{\lambda\to\infty}\frac{\sigma^2}{\lambda}=0.$$

Therefore

$$\lim_{\lambda\to\infty}Var(\hat{\beta}_n(CAR))=0. \quad \square$$

From the above,

$$MSE(\hat{\beta}_n(CAR))=Bias^T Bias+trace[Var(\hat{\beta}_n(CAR))].$$

We have

$$\begin{aligned} Bias^T Bias &= [\lambda(\mathbf{X}^T\mathbf{X}+\lambda W)^{-1}W\beta]^T[\lambda(\mathbf{X}^T\mathbf{X}+\lambda W)^{-1}W\beta] \\ &= \lambda\beta^T W\frac{1}{\lambda}(\frac{\mathbf{X}^T\mathbf{X}}{\lambda}+W)^{-1}\lambda\frac{1}{\lambda}(\frac{\mathbf{X}^T\mathbf{X}}{\lambda}+W)^{-1}W\beta \\ &= \beta^T W(\frac{\mathbf{X}^T\mathbf{X}}{\lambda}+W)^{-1}(\frac{\mathbf{X}^T\mathbf{X}}{\lambda}+W)^{-1}W\beta. \end{aligned}$$

So

$$\lim_{\lambda\to\infty}Bias^T Bias=\beta^T\beta.$$

Recall that the ordinary least square estimator $\hat{\beta}_n(OLS)$ is unbiased and has a variance-covariance matrix $Var(\hat{\beta}_n(OLS))=\sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$. Its mean squared error is

$$\begin{aligned} MSE(\hat{\beta}_n(OLS)) &= 0+trace[Var(\hat{\beta}_n(OLS))] \\ &= trace[\sigma^2(\mathbf{X}^T\mathbf{X})^{-1}]. \end{aligned}$$

For the OSL with serious issue of multicollinearity, entries of $(\mathbf{X}^T\mathbf{X})^{-1}$ are large, so $trace[\sigma^2(\mathbf{X}^T\mathbf{X})^{-1}]$ is large. Although $Bias^T Bias = \beta^T\beta > 0$ for CAR, but $\lim_{\lambda\to\infty} trace[Var(\hat{\beta}_n(CAR))] = 0$. Therefore for large λ, the mean squared error for CAR is likely smaller than the mean squared error for the OLS.

Next, we show that after suitable data argumentation, the CAR regression is equivalent to an OLS regression. The idea is similar to that in Anbari and Mkhadri (2008).

Theorem 4.2.2. Given the Cholesky's decomposition $W = CC^T$ and $\lambda > 0$, define

$$\mathbf{X}^* = \frac{1}{\sqrt{1+\lambda}}\begin{pmatrix} \mathbf{X} \\ \sqrt{\lambda}C^T \end{pmatrix}, \quad \mathbf{Y}^* = \begin{pmatrix} \mathbf{Y} \\ 0 \end{pmatrix}, \quad \beta^* = \sqrt{1+\lambda}\beta.$$

Then minimizing $CARLS$ is equivalent to minimizing the $OLS = \sum_{i=1}^{n+p}(y_i^* - x_i^*\beta^*)^2 = (\mathbf{Y}^* - \mathbf{X}^*\beta^*)^T(\mathbf{Y}^* - \mathbf{X}^*\beta^*)$, where x_i^* is the i^{th} row of $\mathbf{X}^*$.

Proof.

$$\begin{aligned}
OLS &= \sum_{i=1}^{n+p}(y_i^* - x_i^*\beta^*)^2 \\
&= (\mathbf{Y}^* - \mathbf{X}^*\beta^*)^T(\mathbf{Y}^* - \mathbf{X}^*\beta^*) \\
&= [(\mathbf{Y}^*)^T - (\beta^*)^T(\mathbf{X}^*)^T][\mathbf{Y}^* - \mathbf{X}^*\beta^*] \\
&= (\mathbf{Y}^*)^T\mathbf{Y}^* - (\mathbf{Y}^*)^T\mathbf{X}^*\beta^* - (\beta^*)^T(\mathbf{X}^*)^T\mathbf{Y}^* + (\mathbf{X}^*\beta^*)^T(\mathbf{X}^*\beta^*) \\
&= \begin{pmatrix} \mathbf{Y}^T & 0 \end{pmatrix}\begin{pmatrix} \mathbf{Y} \\ 0 \end{pmatrix} - \begin{pmatrix} \mathbf{Y}^T & 0 \end{pmatrix}\frac{1}{\sqrt{1+\lambda}}\begin{pmatrix} \mathbf{X} \\ \sqrt{\lambda}C^T \end{pmatrix}\sqrt{1+\lambda}\beta \\
&\quad -\sqrt{1+\lambda}\beta^T\frac{1}{\sqrt{1+\lambda}}\begin{pmatrix} \mathbf{X}^T\sqrt{\lambda}C \end{pmatrix}\begin{pmatrix} \mathbf{Y} \\ 0 \end{pmatrix} \\
&\quad +\sqrt{1+\lambda}\beta^T\frac{1}{\sqrt{1+\lambda}}\begin{pmatrix} \mathbf{X}^T\sqrt{\lambda}C \end{pmatrix}\frac{1}{\sqrt{1+\lambda}}\begin{pmatrix} \mathbf{X} \\ \sqrt{\lambda}C^T \end{pmatrix}\sqrt{1+\lambda}\beta \\
&= \mathbf{Y}^T\mathbf{Y} - \mathbf{Y}^T\mathbf{X}\beta - \beta^T\mathbf{X}^T\mathbf{Y} + \beta^T(\mathbf{X}^T\mathbf{X} + \lambda CC^T)\beta \\
&= (\mathbf{Y} - \mathbf{X}\beta)^T(\mathbf{Y} - \mathbf{X}\beta) + \lambda\beta^T W\beta = CARLS. \quad \square
\end{aligned}$$

It is important to select the best λ according to certain criterion. There are several criteria available for this purpose, including the Bayesian Information Criterion (BIC), Alkaike Information Criterion (AIC) and Leave-One-Out criterion. We would discuss the Leave-One-Out method here.

If the observation $(Y_i, X_{i1}, X_{i2}, \cdots, X_{ip})$ is removed from the dataset, $i = 1, 2, \cdots, n$, let the $\mathbf{X}$ matrix after deleting the i^{th} row be $\mathbf{X}_{(i)}$ and the $\mathbf{Y}$ matrix after deleting the i^{th} element be $\mathbf{Y}_{(i)}$. For each $i = 1, 2, \cdots, n$, define

$$\hat{\beta}_{(n-1)}(CAR_i) = \left(\mathbf{X}_{(i)}^T\mathbf{X}_{(i)} + \lambda W\right)^{-1}\left(\mathbf{X}_{(i)}^T\mathbf{Y}_{(i)}\right)$$

and the prediction error square for $(Y_i, X_{i1}, X_{i2}, \cdots, X_{ip})$ is

$$CV_{(-i)}(\lambda) = \left(y_i - \sum_{j=1}^{p} x_{ij} \hat{\beta}_{(n-1)}(CAR_i)_j \right)^2 .$$

The CV of λ is

$$CV(\lambda) = \frac{1}{n} \sum_{i=1}^{n} CV_{(-i)}(\lambda).$$

We look for the value of λ that minimizes $CV(\lambda)$.

4.3 CAR for logistic models

We extend the above two types of correlation adjusted penalties to logistic regression. Recall that the log-likelihood function for logistic regression is

$$\ell(\beta) = \sum_{i=1}^{n} [y_i \log(\pi_i) + (1 - y_i) \log(1 - \pi_i)],$$

where $\pi_i = P(Y_i = 1) = \frac{1}{1+e^{-x_i \beta}} = \frac{e^{x_i \beta}}{1+e^{x_i \beta}} = P(x_i)$.

4.3.1 Formulation of the problem

Because minimizing the OLS for the multiple linear regression is equivalent to maximizing the log-likelihood function, we focus on maximizing the log-likelihood function for logistic regression. For the first type correlation adjusted penalty, we maximize

$$CARLR_1 = \ell(\beta) - \lambda \left[\sum_{j=1}^{p-1} (\beta_j - r_{j,j+1} \beta_{j+1})^2 + \beta_p^2 \right] = \ell(\beta) - \lambda \beta^T W_1 \beta,$$

and for the second type of penalty, we maximize

$$CARLR_2 = \ell(\beta) - \lambda \left[\sum_{j=1}^{p-1} \sum_{k>j} (\beta_j - r_{j,k}\beta_k)^2 + \beta_p^2 \right] = \ell(\beta) - \lambda\beta^T W_2 \beta.$$

That is, we maximize

$$CARLR = \ell(\beta) - \lambda\beta^T W \beta$$

where W can be either W_1 or W_2.

4.3.2 Some properties

The ridge penalty for logistic regression has been investigated by several authors including Le Cessie and Van Houwelingen (1992) and Barker and Brown (2001). The ridge logistic regression estimator maximizes

$$\ell^\lambda(\beta) = \ell(\beta) - \lambda\beta^T\beta$$

and is $\hat{\beta}^\lambda(MLE) = (\mathbf{X}^T K\mathbf{X}+\lambda I)^{-1}\mathbf{X}^T K\mathbf{X}\hat{\beta}_n(MLE)$ where $\hat{\beta}_n(MLE)$ is the MLE for maximizing $\ell(\beta)$ and K is the diagonal matrix of the MLE of probabilities $\pi_i(1-\pi_i), i = 1, 2, \cdots, n$. As an extension of the ridge regression, a plug-in estimator for correlation adjusted logistic regression may be defined as

$$\hat{\beta}_n(CARLR) = (\mathbf{X}^T K\mathbf{X} + \lambda W)^{-1}\mathbf{X}^T K\mathbf{X}\hat{\beta}_n(MLE).$$

However the performance of this estimator is unknown.

We extend a key idea and result in Gao and Shen (2007). However the proof in Gao and Shen (2007) contains a minor mistake.

We denote the MLE for the correlation adjusted logistic regression as $\hat{\beta}_n(CARLR)$ and show that it follows asymptotically a normal distribution.

Theorem 4.3.1. Let β_0 be the true unknown parameter of the logistic regression. Under regularity conditions for the likelihood function, the correlation adjusted MLE $\hat{\beta}_n(CARLR)$ is asymptotically normally distributed. That is,

$$\sqrt{n}(\hat{\beta}_n(CARLR) - \beta_0) \xrightarrow{d} N(0, I^{-1}(\beta_0))$$

where $I(\beta_0)$ is the Fisher information matrix for the logistic regression evaluated at β_0.

Proof. Consider the score function $S(\beta) = \frac{\partial(CARLR)}{\partial\beta} = \frac{\partial[\ell(\beta)-\lambda\beta^T W\beta]}{\partial\beta}$. Then the MLE $\hat{\beta}_n(CARLR)$ satisfies $S(\hat{\beta}_n(CARLR)) = 0$. The first order Taylor expansion of $S(\beta)$ at β_0 gives, approximately,

$$\begin{aligned} 0 &= S(\hat{\beta}_n(CARLR)) \\ &= S(\beta_0) + S^{'}(\beta_0)\left(\hat{\beta}_n(CARLS) - \beta_0\right) + o_p\left(||\hat{\beta}_n(CARLS) - \beta_0||\right) \\ &= \left(\frac{\partial\ell(\beta)}{\partial\beta}\Big|_{\beta_0} - 2\lambda W\beta\Big|_{\beta_0}\right) + \left(\frac{\partial^2\ell(\beta)}{\partial\beta^2}\Big|_{\beta_0} - 2\lambda W\right)\left(\hat{\beta}_n(CARLS) - \beta_0\right) \\ &\quad + o_p\left(||\hat{\beta}_n(CARLS) - \beta_0||\right). \end{aligned}$$

Rearranging the terms and removing the higher order terms, we have

$$\begin{aligned} 0 &\approx \left(\frac{\partial\ell(\beta)}{\partial\beta}\Big|_{\beta_0} - 2\lambda W\beta\Big|_{\beta_0}\right) \\ &\quad + \left(\frac{\partial^2\ell(\beta)}{\partial\beta^2}\Big|_{\beta_0} - 2\lambda W\right)\left(\hat{\beta}_n(CARLS) - \beta_0\right). \end{aligned}$$

This gives us

$$\begin{aligned}&\sqrt{n}\left(\hat{\beta}_n(CARLR) - \beta_0\right)\\ \approx\quad & \left[-\frac{1}{n}\left(\frac{\partial^2 \ell(\beta)}{\partial \beta^2}\Big|_{\beta_0} - 2\lambda W\right)\right]^{-1}\left[\frac{1}{\sqrt{n}}\left(\frac{\partial \ell(\beta)}{\partial \beta}\Big|_{\beta_0} - 2\lambda W\beta_0\right)\right].\end{aligned}$$

Because both W and β_0 have bounded elements, we have $\frac{1}{n}(2\lambda W) \xrightarrow{P} 0$ and $\frac{1}{\sqrt{n}}(2\lambda W\beta_0) \xrightarrow{P} 0$. Now

$$\frac{1}{\sqrt{n}}\left(\frac{\partial \ell(\beta)}{\partial \beta}\Big|_{\beta_0}\right) \quad = \quad \sqrt{n}\frac{\sum_{i=1}^{n}\frac{\partial \ln f(x_i,\beta)}{\partial \beta}}{n}.$$

Furthermore,

$$E\left(\frac{\sum_{i=1}^{n}\frac{\partial \ln f(x_i,\beta)}{\partial \beta}}{n}\right) = E\left(\frac{\partial \ln f(x,\beta)}{\partial \beta}\right) = 0,$$

and

$$E\left(\frac{\partial \ln f(x,\beta)}{\partial \beta_i}\frac{\partial \ln f(x,\beta)}{\partial \beta_j}\right)_{i,j} = I(\beta).$$

By Multivariate Central Limit Theorem, as $n \to \infty$,

$$\frac{1}{\sqrt{n}}\left(\frac{\partial \ell(\beta)}{\partial \beta}\Big|_{\beta_0}\right) \xrightarrow{d} N\left(0, I(\beta_0)\right).$$

Now,

$$-\frac{1}{n}\left(\frac{\partial^2 \ell(\beta)}{\partial \beta^2}\right) = \frac{\sum_{i=1}^{n} -\frac{\partial^2 \ln f(x_i,\beta)}{\partial \beta^2}}{n},$$

and

$$E\left(\frac{\sum_{i=1}^{n} -\frac{\partial^2 \ln f(x_i,\beta)}{\partial \beta^2}}{n}\right) = E\left(-\frac{\partial^2 \ln f(x,\beta)}{\partial \beta^2}\right).$$

Since

$$E\left(-\frac{\partial^2 \ln f(x,\beta)}{\partial \beta^2}\right) = I(\beta),$$

by multivariate Law of Large Numbers, as $n \to \infty$,

$$-\frac{1}{n}\frac{\partial^2 \ell(\beta)}{\partial \beta^2}\Big|_{\beta_0} \xrightarrow{P} I(\beta_0).$$

Therefore by Slutsky's Theorem,

$$\left[-\frac{1}{n}\left(\frac{\partial^2 \ell(\beta)}{\partial \beta^2}\Big|_{\beta_0} - 2\lambda W\right)\right]^{-1} \xrightarrow{P} I^{-1}(\beta_0).$$

Apply Slutsky's Theorem again, as $n \to \infty$,

$$\sqrt{n}\left(\hat{\beta}_n(CARLR) - \beta_0\right) \xrightarrow{d} I^{-1}(\beta_0)N(0, I(\beta_0)) = N(0, I^{-1}(\beta_0)). \quad \square$$

4.4 CAR for LAD regression

We now extend CAR to LAD (Least Absolute Deviation) regression. Essentially we extend a result in Xu and Ying (2010) of LASSO-type penalty for LAD.

As indicated by Xu and Ying (2010), the LAD or L_1 method is a good nonlinear alternative to the least squares method and has good robustness properties. The linear regression model is generalized to

$$Y_i = \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} + \epsilon_i, \qquad i = 1, 2, \cdots, n,$$

where the design matrix is known and $\epsilon_i, i = 1, 2, \cdots, n$, are independent and identically distributed random errors with a common distribution F.

The objective of the LAD method is to find the estimator $\hat{\beta}_n(LAD)$ that minimizes

$$LAD(\beta) = \sum_{i=1}^{n} |y_i - (\beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip})|.$$

However there is no explicit form solution for $\hat{\beta}_n(LAD)$. Its derivation is normally carried out by linear programming. Xu and Ying (2010) introduced the penalized LAD

$$LASSOLAD = \frac{LAD(\beta)}{n} + \frac{1}{n}\sum_{j=1}^{p} \lambda_{nj}|\beta_j|$$

and studied the asymptotic behavior of the penalized estimator when $n \to \infty$ and $\frac{\lambda_{nj}}{\sqrt{n}} \to \lambda_{0j} \geq 0$.

In this section, we extend the result in Xu and Ying (2010) to a CAR-type penalty for LAD, defined as

$$CARLAD(\beta) = \frac{LAD(\beta)}{n} + \frac{\lambda_n}{n}\beta^T W \beta,$$

where W, as in previous sections, can be either W_1 or W_2. The objective is to find the estimator $\hat{\beta}_n(CARLAD)$ that minimizes $CARLAD$ (the CAR penalized LAD).

As in Xu and Ying (2010), we make the following two assumptions:

(A.1) The random errors $\epsilon_i, i = 1, 2, \cdots, n$, are independent and identically distributed with median 0 and a density function f which is continuous and strictly positive in a neighborhood of 0;

(A.2) The design matrix $\mathbf{X}$ (depending on n) is deterministic and there is a positive definite matrix Q (of size $p \times p$) such that $\lim_{n\to\infty} \frac{1}{n}\mathbf{X}^T\mathbf{X} = Q^2$.

The following result based on Taylor expansion is from Xu and Ying (2010).

Proposition 4.4.1. Under the above assumptions (A.1) and (A.2), for any sequence $d_n > 0$ such that $d_n \to 0$ in probability, we have

$$\begin{aligned} & \frac{1}{n}[LAD(\beta) - LAD(\beta_0)] \\ = \; & -\frac{1}{n}\sum_{i=1}^{n} sgn(\epsilon_i)x_i(\beta - \beta_0) + \frac{1}{2}f(0)(\beta - \beta_0)^T Q^2(\beta - \beta_0) \\ & +o_p(||\beta - \beta_0||^2 + n^{-1}), \end{aligned}$$

uniformly in $||\beta - \beta_0|| = \sum_{j=1}^{p} |\beta_j - \beta_{0j}| \le d_n$, where β_0 is the true unknown parameter, $x_i = (x_{i1} \; x_{i2} \; \cdots \; x_{ip})$, $sgn(\epsilon_i)$ is the sign of ϵ_i, and $||\beta - \beta_0||^2 = \sum_{j=1}^{p} |\beta_j - \beta_{0j}|^2$.

Xu and Ying (2010) defined the function

$$C(u) = \frac{1}{2}u^T Du - a^T u + \sum_{j=1}^{s} \lambda_j u_j + \sum_{j=s+1}^{p} \lambda_j |u_j|$$

and showed that $C(u) - C(\hat{u}) \ge \frac{1}{2}(u - \hat{u})^T D(u - \hat{u})$ for any $p \times 1$ vectors of real numbers u and $\hat{u}$, where D is a positive definite matrix, a is any $p \times 1$ vector of real numbers, $\lambda_1, \cdots, \lambda_s$ are constants, and $\lambda_{s+1}, \cdots, \lambda_p$ are nonnegative constants. Taking $\lambda_i = 0$ for $i = 1, 2, \cdots, p$, we have

Proposition 4.4.2. For any $p \times 1$ vectors of real numbers u and $\hat{u}$, we have

$$C(u) - C(\hat{u}) \ge \frac{1}{2}(u - \hat{u})^T D(u - \hat{u}),$$

where $C(u) = \frac{1}{2}u^T Du - a^T u$.

To extend the result in Xu and Ying (2010), we now discuss the asymptotic distribution of $\hat{\beta}_n(CARLAD)$.

Theorem 4.4.1. Assume conditions (A.1) and (A.2) and $\lim_{n\to\infty} \frac{\lambda_n}{\sqrt{n}} = \lambda_0 \geq 0$. Then in distribution, as $n \to \infty$,

$$\sqrt{n}(\hat{\beta}_n(CARLAD) - \beta_0) \xrightarrow{d} R$$

where R is the random variable that minimizes

$$R(u) = M^T u + \frac{1}{2} f(0) u^T Q^2 u + 2\lambda_0 (W\beta_0)^T u,$$

and M follows the multivariate normal distribution $N(0, Q^2)$.

Proof. Write $\hat{\beta}_n(CARLAD) = \hat{\beta}_n$. Let $f(\beta) = \beta^T W \beta$, $f^{'}(\beta) = 2W\beta$, $f^{''}(\beta) = 2W$. The Taylor expansion of $\beta^T W \beta$ at β_0 is

$$\begin{aligned} f(\beta) = f(\beta_0) &+ f^{'}(\beta_0)(\beta - \beta_0) \\ &+ \frac{1}{2}(\beta - \beta_0)^T f^{''}(\beta)(\beta - \beta_0) + o_p(||\beta - \beta_0||^2). \end{aligned}$$

Then,

$$\begin{aligned} \hat{\beta}_n^T W \hat{\beta}_n = \beta_0^T W \beta_0 &+ 2(W\beta_0)^T(\hat{\beta}_n - \beta_0) \\ &+ \frac{1}{2}(\hat{\beta}_n - \beta_0)^T 2W(\hat{\beta}_n - \beta_0) + o_p(||\hat{\beta}_n - \beta_0||^2), \end{aligned}$$

so,

$$\begin{aligned} & \hat{\beta}_n^T W \hat{\beta}_n - \beta_0^T W \beta_0 \\ = \ & 2(W\beta_0)^T(\hat{\beta}_n - \beta_0) \\ & + \frac{1}{2}(\hat{\beta}_n - \beta_0)^T 2W(\hat{\beta}_n - \beta_0) + o_p(||\hat{\beta}_n - \beta_0||^2) \\ = \ & 2\beta_0^T W(\hat{\beta}_n - \beta_0) \\ & + (\hat{\beta}_n - \beta_0)^T W(\hat{\beta}_n - \beta_0) + o_p(||\hat{\beta}_n - \beta_0||^2). \end{aligned}$$

Using Proposition 4.4.1 and the Taylor expansion, we have

$$
\begin{aligned}
& CARLAD(\hat{\beta}_n) - CARLAD(\beta_0) \\
= \;& \left[\frac{1}{n}LAD(\hat{\beta}_n) + \frac{\lambda_n}{n}\hat{\beta}_n^T W\hat{\beta}_n\right] - \left[\frac{1}{n}LAD(\hat{\beta}_0) + \frac{\lambda_n}{n}\hat{\beta}_0^T W\hat{\beta}_0\right] \\
= \;& \frac{1}{n}\left[LAD(\hat{\beta}_n) - LAD(\beta_0)\right] + \frac{\lambda_n}{n}\left[\hat{\beta}_n^T W\hat{\beta}_n - \beta_0^T W\beta_0\right] \\
= \;& -\frac{1}{n}\sum_{i=1}^{n} sgn(\epsilon_i)x_i(\hat{\beta}_n - \beta_0) + \frac{1}{2}f(0)(\hat{\beta}_n - \beta_0)^T Q^2(\hat{\beta}_n - \beta_0) \\
& +\frac{\lambda_n}{n}\left[2(W\beta_0)^T(\hat{\beta}_n - \beta_0) + (\hat{\beta}_n - \beta_0)^T W(\hat{\beta}_n - \beta_0)\right] \\
& +o_p(||\hat{\beta}_n - \beta_0||^2 + n^{-1}) + \frac{\lambda_n}{n}o_p(||\hat{\beta}_n - \beta_0||^2) \\
= \;& \frac{1}{n}\{-\sum_{i=1}^{n}\frac{1}{\sqrt{n}}sgn(\epsilon_i)x_i[\sqrt{n}(\hat{\beta}_n - \beta_0)] \\
& +\frac{1}{2}f(0)[\sqrt{n}(\hat{\beta}_n - \beta_0)]^T Q^2[\sqrt{n}(\hat{\beta}_n - \beta_0)] \\
& +2\frac{\lambda_n}{\sqrt{n}}(W\beta_0)^T[\sqrt{n}(\hat{\beta}_n - \beta_0)] + \frac{\lambda_n}{n}[\sqrt{n}(\hat{\beta}_n - \beta_0)]^T W[\sqrt{n}(\hat{\beta}_n - \beta_0)]\} \\
& +o_p(||\hat{\beta}_n - \beta_0||^2 + n^{-1}) + \frac{\lambda_n}{n}o_p(||\hat{\beta}_n - \beta_0||^2).
\end{aligned}
$$

Let $\tilde{u}_n = \sqrt{n}(\hat{\beta}_n - \beta_0)$, then

$$
\begin{aligned}
& CARLAD(\hat{\beta}_n) - CARLAD(\beta_0) \\
= \;& \frac{1}{n}\{-\sum_{i=1}^{n}\frac{1}{\sqrt{n}}sgn(\epsilon_i)x_i\tilde{u}_n + \frac{1}{2}f(0)\tilde{u}_n^T Q^2\tilde{u}_n + 2\frac{\lambda_n}{\sqrt{n}}(W\beta_0)^T\tilde{u}_n + \frac{\lambda_n}{n}\tilde{u}_n^T W\tilde{u}_n\} \\
& +o_p(||\hat{\beta}_n - \beta_0||^2 + n^{-1}) + \frac{\lambda_n}{n}o_p(||\hat{\beta}_n - \beta_0||^2).
\end{aligned}
$$

Set the function

$$
B_n(u) = -\sum_{i=1}^{n}\frac{1}{\sqrt{n}}sgn(\epsilon_i)x_i u + \frac{1}{2}f(0)u^T Q^2 u + 2\frac{\lambda_n}{\sqrt{n}}(W\beta_0)^T u
$$

and let $\hat{u}_n$ be the random vector which minimizes $B_n(u)$. Then

$$B_n(u) = \sqrt{n}\left(\frac{-\sum_{i=1}^n sgn(\epsilon_i)x_i}{n}\right)u + \frac{1}{2}f(0)u^TQ^2u + 2\frac{\lambda_n}{\sqrt{n}}(W\beta_0)^Tu$$

Since $\lim_{n\to\infty}\frac{\lambda_n}{\sqrt{n}} = \lambda_0 \geq 0$, and by Central Limit Theorem, as $n \to \infty$,

$$\sqrt{n}\left(\frac{-\sum_{i=1}^n sgn(\epsilon_i)x_i}{n}\right) \xrightarrow{d} M^T$$

so,

$$B_n(u) \xrightarrow{d} R(u) = M^Tu + \frac{1}{2}f(0)u^TQ^2u + 2\lambda_0(W\beta_0)^Tu.$$

Since $\hat{u}_n$ minimizes $B_n(u)$, and R minimizes $R(u)$, then

$$\hat{u}_n \xrightarrow{d} R$$

and is bounded in probability. Therefore, $\frac{1}{n}\hat{u}_n \xrightarrow{p} 0$ in probability.

Because $\frac{\lambda_n}{n}W \xrightarrow{p} 0$, the 0 matrix, by Slutsky's Theorem, $\tilde{u}_nQ^2\tilde{u}_n + \frac{\lambda_n}{n}\tilde{u}_nW\tilde{u}_n$ is asymptotically equivalent to $\tilde{u}_nQ^2\tilde{u}_n$. Therefore

$$CARLAD(\hat{\beta}_n) - CARLAD(\beta_0) \approx \frac{1}{n}B_n(\tilde{u}_n).$$

We apply the same arguments in Xu and Ying (2010) and conclude that $\tilde{u}_n$ and $\hat{u}_n$ have the same asymptotic distribution. So, $\tilde{u}_n - \hat{u}_n \xrightarrow{p} 0$, and by Slutsky's Theorem,

$$\tilde{u}_n = \sqrt{n}(\hat{\beta}_n - \beta_0) \xrightarrow{d} R \qquad \square$$

4.5 Summary

In this chapter, we introduced a new type of penalization using the sample correlations among the predictor variables. Our motivation was that highly correlated predictor variables should have similar effects on the response variable and so similar parameters. After defining the objective of our penalized regression, we derived the parameter estimator and investigated its properties. For example, we showed that the variance-covariance matrix of the penalized estimator gets smaller when the tuning parameter gets larger, and its asymptotic distribution exists for both the logistic regression and the LAD regression. We also showed that after a suitable data argumentation, the penalized regression becomes the ordinary least squares regression.

Chapter 5

Correlation adjusted elastic net (CAEN)

5.1 Introduction

In this chapter, we extend the two new types of regularization method for simultaneous shrinkage and variable selection introduced in the previous chapter to elastic net and call them Correlation Adjusted Elastic Net (CAEN). They can be regarded as a data-adjusted extension of elastic net regression. Some theoretical results for multiple linear regression, logistic regression and other types of regression are derived. We show that CAEN for multiple linear regression is reduced to LASSO after applying argumentation to the data set, and the parameter estimators for CAEN logistic regression follow asymptotically a normal distribution. Similar results are also derived for the LAD regression.

5.2 CAEN for linear models

We first discuss the two types of CAEN for multiple linear regression. Recall that we assume centered and standardized observations and the OLS regression minimizes

$$OLS = \sum_{i=1}^{n} (y_i - x_i\beta)^2 = (\mathbf{Y} - \mathbf{X}\beta)^T(\mathbf{Y} - \mathbf{X}\beta)$$

where

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}, \quad \mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix},$$

$\beta = (\beta_1, \beta_2, \cdots, \beta_p)^T$ is the vector of unknown regression parameters, and x_i is the i^{th} row of the design matrix $\mathbf{X}$. Recall also that the least squares estimator of β is $\hat{\beta}_n(OLS) = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$ and its variance-covariance matrix is $\sigma^2(\hat{\beta}_n(OLS)) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$ where σ^2 is the population variance of the regression model.

5.2.1 Formulation of the problem

We define two types of CAEN by incorporating empirical correlation coefficients in the penalty function. The first type is defined by CAEN least squares

$$CAEN_1 = OLS + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \left[\sum_{j=1}^{p-1} (\beta_j - r_{j,j+1}\beta_{j+1})^2 + \beta_p^2 \right]$$

where $r_{j,k}$ is the sample correlation between the predictors X_j and X_k. The objective is to find $\hat{\beta}_n(CAEN_1)$ that minimizes $CAEN_1$.

For the first type, define the matrix

$$D_1 = \begin{pmatrix} 1 & -r_{1,2} & 0 & \cdots & 0 & 0 \\ 0 & 1 & -r_{2,3} & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & -r_{p-1,p} \\ 0 & 0 & 0 & \cdots & 0 & 1 \end{pmatrix}.$$

Then $\sum_{j=1}^{p-1}(\beta_j - r_{j,j+1}\beta_{j+1})^2 + \beta_p^2 = \beta^T D_1^T D_1 \beta = \beta^T W_1 \beta$, where $W_1 = D_1^T D_1$. Clearly W_1 is a real symmetric $p \times p$ matrix and positive semi-definite because $\beta^T W_1 \beta \geq 0$ for any vector β. Therefore W_1 admits a Cholesky's decomposition $W_1 = C_1 C_1^T$, where C_1 is an upper triangular matrix.

Moreover if $r_{j,j+1} = 0$ for all $j = 1, 2, \cdots, p-1$, then our first type correlation adjusted regression becomes elastic net and hence our definition includes elastic net as a special case.

The second type is defined by the CAEN least squares

$$CAEN_2 = OLS + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \left[\sum_{j=1}^{p-1} \sum_{k>j} (\beta_j - r_{j,k}\beta_k)^2 + \beta_p^2 \right].$$

The objective is to find $\hat{\beta}_n(CAEN_2)$ that minimizes $CAEN_2$.

For the second type,

$$D_2 = \begin{pmatrix} 1 & -r_{1,2} & 0 & 0 & \cdots & 0 & 0 \\ 1 & 0 & -r_{1,3} & 0 & \cdots & 0 & 0 \\ 1 & 0 & 0 & -r_{1,4} & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & 0 & \cdots & 0 & -r_{1,p} \\ 0 & 1 & -r_{2,3} & 0 & \cdots & 0 & 0 \\ 0 & 1 & 0 & -r_{2,4} & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 1 & 0 & 0 & \cdots & 0 & -r_{2,p} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 1 & -r_{p-1,p} \\ 0 & 0 & 0 & 0 & \cdots & 0 & 1 \end{pmatrix}.$$

Then $\sum_{j=1}^{p-1}\sum_{k>j}(\beta_j - r_{j,k}\beta_k)^2 + \beta_p^2 = \beta^T D_2^T D_2 \beta = \beta^T W_2 \beta$, where $W_2 = D_2^T D_2$. Again W_2 is a real symmetric $p \times p$ matrix and positive semi-definite because $\beta^T W_2 \beta \geq 0$ for any vector β. Therefore again W_2 admits a Cholesky's decomposition $W_2 = C_2 C_2^T$, where C_2 is an upper triangular matrix.

Putting both types together, we minimize

$$CAEN = OLS + \lambda_1 V^T \beta + \lambda_2 \beta^T W \beta$$

where W is either W_1 or W_2 and V is a column vector with i^{th} element being 1 if the sign of the corresponding parameter β_i is positive and -1 if β_i is negative. Although V depends on the unknown parameters thorough their signs, in practice based on theory or empirical evidence, we might be able to determine the signs of the unknown parameters in advance and so V could be regarded as being pre-determined.

The derivative of $CAEN$ with respect to β is given as the column vector

$$\begin{aligned}\frac{d(CAEN)}{d\beta} &= -2[\mathbf{X}^T\mathbf{Y} - (\mathbf{X}^T\mathbf{X})\beta] + \lambda_1 V + 2\lambda_2 W\beta \\ &= -2\left[\mathbf{X}^T\mathbf{Y} - \frac{\lambda_1}{2}V - (\mathbf{X}^T\mathbf{X} + \lambda_2 W)\beta\right].\end{aligned}$$

Therefore the penalized estimator of β is

$$\hat{\beta}_n(CAEN) = (\mathbf{X}^T\mathbf{X} + \lambda_2 W)^{-1}\left(\mathbf{X}^T\mathbf{Y} - \frac{\lambda_1}{2}V\right).$$

This result shows that the MLE of the least squares with a CAEN penalty exists.

5.2.2 Some properties

Comparing with $\hat{\beta}_n(LASSO) = (\mathbf{X}^T\mathbf{X})^{-1}\left(\mathbf{X}^T\mathbf{Y} - \frac{\lambda_1}{2}V\right)$ and $\hat{\beta}_n(ENLS) = (\mathbf{X}^T\mathbf{X} + \lambda_2 I)^{-1}\left(\mathbf{X}^T\mathbf{Y} - \frac{\lambda_1}{2}V\right)$, we see that CAEN estimator is an extension of both LASSO and elastic net estimators by adding the term $\lambda_2 W$ to the matrix $\mathbf{X}^T\mathbf{X}$. Moreover, we can write

$$\begin{aligned}\hat{\beta}_n(CAEN) &= (\mathbf{X}^T\mathbf{X} + \lambda_2 W)^{-1}\mathbf{X}^T\mathbf{Y} - \frac{\lambda_1}{2}(\mathbf{X}^T\mathbf{X} + \lambda_2 W)^{-1}V \\ &= \hat{\beta}_n(CAR) - \frac{\lambda_1}{2}(\mathbf{X}^T\mathbf{X} + \lambda_2 W)^{-1}V.\end{aligned}$$

Because the matrix $\frac{\lambda_1}{2}(\mathbf{X}^T\mathbf{X} + \lambda_2 W)^{-1}V$ does not involve random variables, its variance-covariance matrix is the 0 matrix.

Theorem 5.2.1. For the estimator $\hat{\beta}_n(CAEN)$, we have $\lim\limits_{\lambda_2\to\infty} Var(\hat{\beta}_n(CAEN)) = 0$, the 0 matrix.

Proof. From Theorem 4.2.1,

$$\lim_{\lambda_2\to\infty} Var(\hat{\beta}_n(CAEN)) = \lim_{\lambda_2\to\infty} Var(\hat{\beta}_n(CAR)) = 0. \quad \square$$

From the above,

$$MSE(\hat{\beta}_n(CAEN)) = Bias^T Bias + trace[Var(\hat{\beta}_n(CAEN))].$$

We have,

$$\begin{aligned} Bias &= \lambda_2(\mathbf{X}^T\mathbf{X} + \lambda_2 W)^{-1}W\beta - \frac{\lambda_1}{2}(\mathbf{X}^T\mathbf{X} + \lambda_2 W)^{-1}V \\ &= -(\lambda_2 + \frac{\lambda_1}{2})(\mathbf{X}^T\mathbf{X} + \lambda_2 W)^{-1}(W\beta + V). \end{aligned}$$

Then,

$$\begin{aligned} & Bias^T Bias \\ = \;& (\lambda_2 + \frac{\lambda_1}{2})^2(W\beta+V)^T(\mathbf{X}^T\mathbf{X} + \lambda_2 W)^{-1}(\mathbf{X}^T\mathbf{X} + \lambda_2 W)^{-1}(W\beta+V) \\ = \;& \lambda_2^2(1+\frac{\lambda_1}{\lambda_2})^2(W\beta+V)^T\frac{1}{\lambda_2}(\frac{\mathbf{X}^T\mathbf{X}}{\lambda_2} + W)^{-1}\frac{1}{\lambda_2}(\frac{\mathbf{X}^T\mathbf{X}}{\lambda_2} + W)^{-1}(W\beta+V) \\ = \;& (1+\frac{\lambda_1}{\lambda_2})^2(W\beta+V)^T(\frac{\mathbf{X}^T\mathbf{X}}{\lambda_2} + W)^{-1}(\frac{\mathbf{X}^T\mathbf{X}}{\lambda_2} + W)^{-1}(W\beta+V). \end{aligned}$$

So,

$$\begin{aligned} \lim_{\lambda_2\to\infty} Bias^T Bias &= (W\beta+V)^T W^{-1}W^{-1}(W\beta+V) \\ &= (\beta^T W + V^T)W^{-1}W^{-1}(W\beta+V) \\ &= (\beta + W^{-1}V)^T(\beta + W^{-1}V). \end{aligned}$$

Although $Bias^T Bias > 0$ for CAEN, $\lim_{\lambda_2\to\infty} trace[Var(\hat{\beta}_n(CAEN))] = 0$. Therefore for large λ_2, the MSE of CAEN is likely smaller than that of OLS with serious issue of multicollinearity.

After suitable data argumentation, we show that the CAEN regression is equivalent to a LASSO regression.

Theorem 5.2.2. Given the Cholesky's decomposition $W = CC^T$ and $\lambda_1, \lambda_2 > 0$, define

$$\mathbf{X}^* = \frac{1}{\sqrt{1+\lambda_2}} \begin{pmatrix} \mathbf{X} \\ \sqrt{\lambda_2} C^T \end{pmatrix}, \ \mathbf{Y}^* = \begin{pmatrix} \mathbf{Y} \\ 0 \end{pmatrix}, \ \beta^* = \sqrt{1+\lambda_2}\beta, \ \gamma = \frac{\lambda_1}{\sqrt{1+\lambda_2}}.$$

Then minimizing

$$CAEN = (\mathbf{Y} - \mathbf{X}\beta)^T(\mathbf{Y} - \mathbf{X}\beta) + \lambda_1 V^T \beta + \lambda_2 \beta^T W \beta$$

is equivalent to minimizing

$$LASSO = (\mathbf{Y}^* - \mathbf{X}^*\beta^*)^T(\mathbf{Y}^* - \mathbf{X}^*\beta^*) + \gamma \sum_{j=1}^{p} |\beta_j^*|.$$

Proof. We have

$$\begin{aligned} OLS &= (\mathbf{Y}^* - \mathbf{X}^*\beta^*)^T(\mathbf{Y}^* - \mathbf{X}^*\beta^*) \\ &= (\mathbf{Y}^*)^T\mathbf{Y}^* - (\beta^*)^T(\mathbf{X}^*)^T\mathbf{Y}^* - (\mathbf{Y}^*)^T\mathbf{X}^*\beta^* + (\beta^*)^T(\mathbf{X}^*)^T\mathbf{X}^*\beta^*. \end{aligned}$$

Now,

$$\begin{aligned} (\beta^*)^T(\mathbf{X}^*)^T\mathbf{Y}^* &= (\beta^T\mathbf{X}^T \ \ \sqrt{\lambda_2}\beta^T C) \begin{pmatrix} \mathbf{Y} \\ 0 \end{pmatrix} = \beta^T\mathbf{X}^T\mathbf{Y}, \\ (\mathbf{Y}^*)^T\mathbf{X}^*\beta^* &= (\mathbf{Y}^T \ \ 0) \frac{1}{\sqrt{1+\lambda_2}} \begin{pmatrix} \mathbf{X} \\ \sqrt{\lambda_2} C^T \end{pmatrix} \sqrt{1+\lambda_2}\beta = \mathbf{Y}^T\mathbf{X}\beta, \\ (\beta^*)^T(\mathbf{X}^*)^T\mathbf{X}^*\beta^* &= (\beta^T\mathbf{X}^T \ \ \sqrt{\lambda_2}\beta^T C) \begin{pmatrix} \mathbf{X}\beta \\ \sqrt{\lambda_2} C^T \beta \end{pmatrix} \\ &= \beta^T\mathbf{X}^T\mathbf{X}\beta + \lambda_2 \beta^T C C^T \beta \\ &= \beta^T\mathbf{X}^T\mathbf{X}\beta + \lambda_2 \beta^T W \beta. \end{aligned}$$

Finally,

$$\gamma \sum_{j=1}^{p} |\beta_j^*| = \frac{\lambda_1}{\sqrt{1+\lambda_2}} \sum_{j=1}^{p} |\sqrt{1+\lambda_2}\beta_j| = \lambda_1 \sum_{j=1}^{p} |\beta_j| = \lambda_1 V^T \beta.$$

Therefore,

$$\begin{aligned} OLS &= \mathbf{Y}^T\mathbf{Y} - \beta^T\mathbf{X}^T\mathbf{Y} - \mathbf{Y}^T\mathbf{X}\beta + \beta^T\mathbf{X}^T\mathbf{X}\beta + \lambda_2\beta^T W\beta + \lambda_1 V^T\beta \\ &= (\mathbf{Y}-\mathbf{X}\beta)^T(\mathbf{Y}-\mathbf{X}\beta) + \lambda_2\beta^T W\beta + \lambda_1 V^T\beta = CAEN. \quad \square \end{aligned}$$

5.3 CAEN for logistic models

We extend the two types of correlation adjusted elastic net to logistic regression. Recall that the log-likelihood function for logistic regression is

$$\ell(\beta) = \sum_{i=1}^{n} [y_i \log(\pi_i) + (1-y_i)\log(1-\pi_i)]$$

where $\pi_i = P(Y_i = 1) = \frac{1}{1+e^{-x_i\beta}}$.

5.3.1 Formulation of the problem

Because minimizing the OLS for the multiple linear regression is equivalent to maximizing the log-likelihood function, we focus on maximizing the log-likelihood function for logistic regression. For the first type CAEN, we maximize

$$\begin{aligned} CAENLR_1 &= \ell(\beta) - \lambda_1 \sum_{j=1}^{p} |\beta_j| - \lambda_2 \left[\sum_{j=1}^{p-1} (\beta_j - r_{j,j+1}\beta_{j+1})^2 + \beta_p^2 \right] \\ &= \ell(\beta) - \lambda_1 V^T\beta - \lambda_2 \beta^T W_1 \beta, \end{aligned}$$

and for the second type of penalty, we maximize

$$\begin{aligned} CAENLR_2 &= \ell(\beta) - \lambda_1 \sum_{j=1}^{p} |\beta_j| - \lambda_2 \left[\sum_{j=1}^{p-1} \sum_{k>j} (\beta_j - r_{j,k}\beta_k)^2 + \beta_p^2 \right] \\ &= \ell(\beta) - \lambda_1 V^T \beta - \lambda_2 \beta^T W_2 \beta. \end{aligned}$$

That is, we maximize

$$CAENLR = \ell(\beta) - \lambda_1 V^T \beta - \lambda_2 \beta^T W \beta$$

where W can be either W_1 or W_2 and V is defined in Section 3.5.

5.3.2 Some properties

We denote the MLE for CAEN for logistic regression as $\hat{\beta}_n(CAENLR)$ and show that it follows asymptotically a normal distribution.

Theorem 5.3.1. Let β_0 be the true unknown parameter for the logistic regression. Under regularity conditions for the likelihood function, the CAEN MLE $\hat{\beta}_n(CAENLR)$ is asymptotically normally distributed. That is, in distribution,

$$\sqrt{n}(\hat{\beta}_n(CAENLR) - \beta_0) \xrightarrow{d} N(0, I^{-1}(\beta_0))$$

where $I(\beta_0)$ is the Fisher information matrix for the logistic regression evaluated at β_0.

Proof. Consider the score function $S(\beta) = \frac{\partial(CAENLR)}{\partial\beta}$. Then the MLE $\hat{\beta}_n(CAENLR)$ satisfies $S(\hat{\beta}_n(CAENLR)) = 0$. The first order Taylor expansion of $S(\beta)$ at β_0

gives

$$\begin{aligned}
0 &= S(\hat{\beta}_n(CAENLR)) \\
&= S(\beta_0) + S^{'}(\beta_0)\left(\hat{\beta}_n(CAENLS) - \beta_0\right) + o_p\left(||\hat{\beta}_n(CAENLS) - \beta_0||\right) \\
&= \left(\frac{\partial \ell(\beta)}{\partial \beta}\Big|_{\beta_0} - \lambda_1 V - 2\lambda_2 W\beta\Big|_{\beta_0}\right) \\
&\quad + \left(\frac{\partial^2 \ell(\beta)}{\partial \beta^2}\Big|_{\beta_0} - 2\lambda_2 W\right)\left(\hat{\beta}_n(CAENLR) - \beta_0\right) \\
&\quad + o_p\left(||\hat{\beta}_n(CAENLS) - \beta_0||\right).
\end{aligned}$$

Rearranging the terms and removing the higher order terms, we have

$$\begin{aligned}
&\sqrt{n}\left(\hat{\beta}_n(CAENLR) - \beta_0\right) \\
\approx\ & \left[-\frac{1}{n}\left(\frac{\partial^2 \ell(\beta)}{\partial \beta^2}\Big|_{\beta_0} - 2\lambda_2 W\right)\right]^{-1}\left[\frac{1}{\sqrt{n}}\left(\frac{\partial \ell(\beta)}{\partial \beta}\Big|_{\beta_0} - \lambda_1 V - 2\lambda_2 W\beta_0\right)\right].
\end{aligned}$$

Since V, W and β_0 all have bounded elements, we have $\frac{1}{\sqrt{n}}(\lambda_1 V) \xrightarrow{P} 0$, $\frac{1}{\sqrt{n}}(2\lambda_2 W) \xrightarrow{P} 0$ and $\frac{1}{\sqrt{n}}(2\lambda_2 W\beta_0) \xrightarrow{P} 0$. Now

$$\frac{1}{\sqrt{n}}\left(\frac{\partial \ell(\beta)}{\partial \beta}\Big|_{\beta_0}\right) = \sqrt{n}\frac{\sum_{i=1}^n \frac{\partial \ln f(x_i,\beta)}{\partial \beta}}{n}.$$

Since

$$E\left(\frac{\sum_{i=1}^n \frac{\partial \ln f(x_i,\beta)}{\partial \beta}}{n}\right) = E\left(\frac{\partial \ln f(x,\beta)}{\partial \beta}\right) = 0$$

and

$$E\left(\frac{\partial \ln f(x,\beta)}{\partial \beta_i}\frac{\partial \ln f(x,\beta)}{\partial \beta_j}\right)_{i,j} = I(\beta),$$

by the Multivariate Central Limit Theorem, as $n \to \infty$,

$$\frac{1}{\sqrt{n}}\left(\frac{\partial \ell(\beta)}{\partial \beta}\Big|_{\beta_0}\right) \xrightarrow{d} N(0, I(\beta_0)).$$

Now

$$-\frac{1}{n}\left(\frac{\partial^2 \ell(\beta)}{\partial \beta^2}\right) = \frac{\sum_{i=1}^{n} -\frac{\partial^2 \ln f(x_i,\beta)}{\partial \beta^2}}{n}.$$

Since

$$E\left(\frac{\sum_{i=1}^{n} -\frac{\partial^2 \ln f(x_i,\beta)}{\partial \beta^2}}{n}\right) = E\left(-\frac{\partial^2 \ln f(x,\beta)}{\partial \beta^2}\right) = I(\beta),$$

by the multivariate Law of Large Numbers, as $n \to \infty$,

$$-\frac{1}{n}\frac{\partial^2 \ell(\beta)}{\partial \beta^2}\Big|_{\beta_0} \xrightarrow{P} I(\beta_0).$$

Therefore by Slutsky's Theorem,

$$\left[-\frac{1}{n}\left(\frac{\partial^2 \ell(\beta)}{\partial \beta^2}\Big|_{\beta_0} - 2\lambda_2 W\right)\right]^{-1} \xrightarrow{P} I^{-1}(\beta_0).$$

Apply Slutsky's Theorem again, as $n \to \infty$.

$$\sqrt{n}\left(\hat{\beta}_n(CAENLR) - \beta_0\right) \xrightarrow{d} I^{-1}(\beta_0)N(0, I(\beta_0)) = N(0, I^{-1}(\beta_0)). \qquad \square$$

5.4 CAEN for LAD regression

Using similar ideas in Section 4.4, we extend CAEN to LAD (Least Absolute Deviation) regression. In fact this is easier because some results in Xu and Ying (2010) of LASSO-type penalty for LAD could be directly used.

As indicated by Xu and Ying (2010), the LAD or L_1 method is a good nonlinear alternative to the least squares method and has good robustness properties. The linear regression model is generalized to

$$Y_i = \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} + \epsilon_i, \qquad i = 1, 2, \cdots, n,$$

where the design matrix is known and $\epsilon_i, i = 1, 2, \cdots, n$, are independent and identically distributed random errors with a common distribution F.

The objective of the LAD method is to find the estimator $\hat{\beta}_n(LAD)$ that minimizes

$$LAD(\beta) = \sum_{i=1}^{n} |y_i - (\beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip})|.$$

However there is no explicit form solution for $\hat{\beta}_n(LAD)$. Its derivation is normally carried out by linear programming. Xu and Ying (2010) introduced the penalized LAD

$$LASSOLAD = \frac{LAD(\beta)}{n} + \frac{1}{n}\sum_{j=1}^{p} \lambda_{nj}|\beta_j|$$

and studied the asymptotic behavior of the penalized estimator when $n \to \infty$ and $\frac{\lambda_{nj}}{\sqrt{n}} \to \lambda_{0j} \geq 0$.

In this section, we extend the result in Xu and Ying (2010) to our CAEN-type penalty for LAD regression, defined as

$$CAENLAD(\beta) = \frac{LAD(\beta)}{n} + \frac{1}{n}\sum_{j=1}^{p} \lambda_{nj}|\beta_j| + \frac{\lambda_n^*}{n}\beta^T W \beta,$$

where W is as in previous sections and can be either W_1 or W_2. The objective is to find the estimator $\hat{\beta}_n(CAENLAD)$ that minimizes $CAENLAD$, the CAEN penalized LAD.

As in Xu and Ying (2010), we make the following two assumptions:

(A.1) The random errors $\epsilon_i, i = 1, 2, \cdots, n$, are independent and identically distributed with median 0 and a density function f which is continuous and strictly positive in a neighborhood of 0;

(A.2) The design matrix $\mathbf{X}$ (depending on n) is deterministic and there is a positive definite matrix Q (of size $p \times p$) such that $\lim_{n\to\infty} \frac{1}{n}\mathbf{X}^T\mathbf{X} = Q^2$.

The following result is from Xu and Ying (2010) and is based on Taylor expansion.

Proposition 5.4.1. Under the above assumptions (A.1) and (A.2), for any sequence $d_n > 0$ such that $d_n \to 0$ in probability, we have

$$\begin{aligned} & \frac{1}{n}[LAD(\beta) - LAD(\beta_0)] \\ = \; & -\frac{1}{n}\sum_{i=1}^{n} sgn(\epsilon_i)x_i(\beta - \beta_0) \\ & +\frac{1}{2}f(0)(\beta-\beta_0)^T Q^2(\beta-\beta_0) + o_p(||\beta-\beta_0||^2 + n^{-1}), \end{aligned}$$

uniformly in $||\beta - \beta_0|| = \sum_{j=1}^{p} |\beta_j - \beta_{0j}| \leq d_n$, where β_0 is the true unknown parameter, $x_i = (x_{i1} \; x_{i2} \; \cdots \; x_{ip})$, $sgn(\epsilon_i)$ is the sign of ϵ_i, and $||\beta - \beta_0||^2 = \sum_{j=1}^{p} |\beta_j - \beta_{0j}|^2$.

Xu and Ying (2010) defined the function

$$C(u) = \frac{1}{2}u^T Du - a^T u + \sum_{j=1}^{s} \lambda_j u_j + \sum_{j=s+1}^{p} \lambda_j |u_j|,$$

where D is a positive definite matrix, a is any $p \times 1$ vector of real numbers, $\lambda_1, \cdots, \lambda_s$ are constants, and $\lambda_{s+1}, \cdots, \lambda_p$ are nonnegative constants. Xu and Ying (2010) showed that

Proposition 5.4.2. For any $p \times 1$ vectors of real numbers u and $\hat{u}$, we have

$$C(u) - C(\hat{u}) \geq \frac{1}{2}(u - \hat{u})^T D(u - \hat{u}),$$

where $C(u) = \frac{1}{2}u^T Du - a^T u + \sum_{j=1}^{s} \lambda_j u_j + \sum_{j=s+1}^{p} \lambda_j |u_j|$.

To extend the result in Xu and Ying (2010), we now discuss the asymptotic distribution of $\hat{\beta}_n(CAENLAD)$.

Theorem 5.4.1. Assume conditions (A.1) and (A.2), $\lim_{n\to\infty} \frac{\lambda_{nj}}{\sqrt{n}} = \lambda_{0j} \geq 0$ and $\lim_{n\to\infty} \frac{\lambda_n^*}{\sqrt{n}} = \lambda_0^* \geq 0$. Then in distribution, as $n \to \infty$,

$$\sqrt{n}(\hat{\beta}_n(CAENLAD) - \beta_0) \xrightarrow{d} R$$

where R is the random variable that minimizes

$$R(u) = M^T u + \frac{f(0)}{2} u^T Q^2 u + \sum_{j=1}^{p} \lambda_{0j} sgn(\beta_{0j}) u_j + 2\lambda_0^* (W\beta_0)^T u,$$

and M follows the multivariate normal distribution $N(0, Q^2)$.

Proof. Write $\hat{\beta}_n(CAENLAD) = \hat{\beta}_n$. Let $f(\beta) = \beta^T W \beta$, $f^{'}(\beta) = 2W\beta$, $f^{''}(\beta) = 2W$. The Taylor expansion of $\beta^T W \beta$ at β_0 is

$$\begin{aligned} f(\beta) = f(\beta_0) &+ f^{'}(\beta_0)(\beta - \beta_0) \\ &+ \frac{1}{2}(\beta - \beta_0)^T f^{''}(\beta)(\beta - \beta_0) + o_p(||\beta - \beta_0||^2). \end{aligned}$$

Then,

$$\begin{aligned} \hat{\beta}_n^T W \hat{\beta}_n = \beta_0^T W \beta_0 &+ 2(W\beta_0)^T(\hat{\beta}_n - \beta_0) \\ &+ \frac{1}{2}(\hat{\beta}_n - \beta_0)^T 2W(\hat{\beta}_n - \beta_0) + o_p(||\hat{\beta}_n - \beta_0||^2) \end{aligned}$$

so,

$$
\begin{aligned}
& \hat{\beta}_n^T W \hat{\beta}_n - \beta_0^T W \beta_0 \\
= \; & 2(W\beta_0)^T(\hat{\beta}_n - \beta_0) \\
& + \frac{1}{2}(\hat{\beta}_n - \beta_0)^T 2W(\hat{\beta}_n - \beta_0) + o_p(||\hat{\beta}_n - \beta_0||^2) \\
= \; & 2\beta_0^T W(\hat{\beta}_n - \beta_0) \\
& + (\hat{\beta}_n - \beta_0)^T W(\hat{\beta}_n - \beta_0) + o_p(||\hat{\beta}_n - \beta_0||^2).
\end{aligned}
$$

Let $\hat{\beta}_n^j$ be the j^{th} component of $\hat{\beta}_n$.

$$
\begin{aligned}
& CAENLAD(\hat{\beta}_n) - CAENLAD(\beta_0) \\
= \; & \left(\frac{1}{n}LAD(\hat{\beta}_n) + \frac{1}{n}\sum_{j=1}^{p}\lambda_{nj}|\hat{\beta}_n^j| + \frac{\lambda_n^*}{n}\hat{\beta}_n^T W \hat{\beta}_n\right) \\
& - \left(\frac{1}{n}LAD(\hat{\beta}_0) + \frac{1}{n}\sum_{j=1}^{p}\lambda_{nj}|\hat{\beta}_0^j| + \frac{\lambda_n^*}{n}\hat{\beta}_0^T W \hat{\beta}_0\right) \\
= \; & \frac{1}{n}\left(LAD(\hat{\beta}_n) - LAD(\hat{\beta}_0)\right) + \frac{1}{n}\left(\sum_{j=1}^{p}\lambda_{nj}|\hat{\beta}_n^j| - \sum_{j=1}^{p}\lambda_{nj}|\hat{\beta}_0^j|\right) \\
& + \frac{\lambda_n^*}{n}\left(\hat{\beta}_n^T W \hat{\beta}_n - \hat{\beta}_0^T W \hat{\beta}_0\right)
\end{aligned}
$$

Using Proposition 5.4.1 and the above Taylor expansion,

$$
\begin{aligned}
& CAENLAD(\hat{\beta}_n) - CAENLAD(\beta_0) \\
= & -\frac{1}{n}\sum_{i=1}^{n} sgn(\epsilon_i)x_i(\hat{\beta}_n - \beta_0) + \frac{f(0)}{2}(\hat{\beta}_n - \beta_0)^T Q^2 (\hat{\beta}_n - \beta_0) \\
& +\frac{1}{n}\left(\sum_{j=1}^{P} \frac{\lambda_{nj}}{\sqrt{n}} sgn(\beta_{0j})(\hat{\beta}_n^j - \beta_{0j})\right) \\
& +\frac{\lambda_n^*}{n}\left(2(W\beta_0)^T(\hat{\beta}_n - \beta_0) + (\hat{\beta}_n - \beta_0)^T W (\hat{\beta}_n - \beta_0)\right) \\
& +o_p(||\hat{\beta}_n - \beta_0||^2 + n^{-1}) + \frac{\lambda_n^*}{n} o_p(||\hat{\beta}_n - \beta_0||^2) \\
= & \frac{1}{n}\{-\sum_{i=1}^{n} \frac{1}{\sqrt{n}} sgn(\epsilon_i)x_i[\sqrt{n}(\hat{\beta}_n - \beta_0)] + \frac{f(0)}{2}[\sqrt{n}(\hat{\beta}_n - \beta_0)]^T Q^2 [\sqrt{n}(\hat{\beta}_n - \beta_0)] \\
& +\sum_{j=1}^{P} \frac{\lambda_{nj}}{\sqrt{n}} sgn(\beta_{0j})[\sqrt{n}(\hat{\beta}_n^j - \beta_{0j})] \\
& +2\frac{\lambda_n^*}{\sqrt{n}}(W\beta_0)^T[\sqrt{n}(\hat{\beta}_n - \beta_0)] + \frac{\lambda_n^*}{n}[\sqrt{n}(\hat{\beta}_n - \beta_0)]^T W [\sqrt{n}(\hat{\beta}_n - \beta_0)]\} \\
& +o_p(||\hat{\beta}_n - \beta_0||^2 + n^{-1}) + \frac{\lambda_n^*}{n} o_p(||\hat{\beta}_n - \beta_0||^2).
\end{aligned}
$$

Define $\tilde{u}_n = \sqrt{n}(\hat{\beta}_n - \beta_0)$ and $\tilde{u}_n^j = \sqrt{n}(\hat{\beta}_n^j - \beta_0^j)$, then

$$
\begin{aligned}
& CAENLAD(\hat{\beta}_n) - CAENLAD(\beta_0) \\
= & \frac{1}{n}\{-\sum_{i=1}^{n} \frac{1}{\sqrt{n}} sgn(\epsilon_i)x_i\tilde{u}_n + \frac{f(0)}{2}\tilde{u}_n^T Q^2 \tilde{u}_n + \sum_{j=1}^{P} \frac{\lambda_{nj}}{\sqrt{n}} sgn(\beta_{0j})\tilde{u}_n^j \\
& +2\frac{\lambda_n^*}{\sqrt{n}}(W\beta_0)^T\tilde{u}_n + \frac{\lambda_n^*}{n}\tilde{u}_n^T W \tilde{u}_n\} \\
& +o_p(||\hat{\beta}_n - \beta_0||^2 + n^{-1}) + \frac{\lambda_n^*}{n} o_p(||\hat{\beta}_n - \beta_0||^2).
\end{aligned}
$$

Set the function

$$\begin{aligned} B_n(u) &= -\sum_{i=1}^{n} \frac{1}{\sqrt{n}} sgn(\epsilon_i) x_i u + \frac{f(0)}{2} u^T Q^2 u + \sum_{j=1}^{p} \frac{\lambda_{nj}}{\sqrt{n}} sgn(\beta_{0j}) u_j \\ &+2 \frac{\lambda_n^*}{\sqrt{n}} (W\beta_0)^T u \end{aligned}$$

and let $\hat{u}_n$ be the random vector which minimizes $B_n(u)$.

$$\begin{aligned} B_n(u) &= \sqrt{n} \left(\frac{-\sum_{i=1}^{n} sgn(\epsilon_i) x_i}{n} \right) u + \frac{f(0)}{2} u^T Q^2 u + \sum_{j=1}^{p} \frac{\lambda_{nj}}{\sqrt{n}} sgn(\beta_{0j}) u_j \\ &+2 \frac{\lambda_n^*}{\sqrt{n}} (W\beta_0)^T u \end{aligned}$$

Since $\lim_{n\to\infty} \frac{\lambda_{nj}}{\sqrt{n}} = \lambda_{0j} \geq 0$ and $\lim_{n\to\infty} \frac{\lambda_n^*}{\sqrt{n}} = \lambda_0^* \geq 0$, by Central Limit Theorem, as $n \to \infty$,

$$\sqrt{n} \left(\frac{-\sum_{i=1}^{n} sgn(\epsilon_i) x_i}{n} \right) \xrightarrow{d} M^T$$

so,

$$B_n(u) \xrightarrow{d} R(u) = M^T u + \frac{f(0)}{2} u^T Q^2 u + \sum_{j=1}^{p} \lambda_{0j} sgn(\beta_{0j}) u_j + 2\lambda_0^* (W\beta_0)^T u.$$

Since $\hat{u}_n$ minimizes $B_n(u)$, and R minimizes $R(u)$, then

$$\hat{u}_n \xrightarrow{d} R$$

and is bounded in probability. Therefore, $\frac{1}{n}\hat{u}_n \xrightarrow{p} 0$ in probability.

Because $\frac{\lambda_n^*}{n} W \xrightarrow{d} 0$, the 0 matrix, by Slutsky's Theorem, $\tilde{u}_n Q^2 \tilde{u}_n + \frac{\lambda_n^*}{n} \tilde{u}_n W \tilde{u}_n$ is asymptotically equivalent to $\tilde{u}_n Q^2 \tilde{u}_n$. Therefore

$$CAENLAD(\hat{\beta}_n) - CAENLAD(\beta_0) \approx \frac{1}{n} B_n(\tilde{u}_n).$$

We apply the same arguments in Xu and Ying (2010) and conclude that $\tilde{u}_n$ and $\hat{u}_n$ have the same asymptotic distribution. so, $\tilde{u}_n - \hat{u}_n \xrightarrow{p} 0$, by Slutsky's Theorem,

$$\tilde{u}_n = \sqrt{n}(\hat{\beta}_n - \beta_0) \xrightarrow{d} R \qquad \square$$

5.5 Summary

In this chapter, We proposed a new type of penalization using the sample correlations among the predictor variables and incorporate them with elastic net. Our motivation is that highly correlated predictor variables should have similar effects on the response variable and so similar parameters. After defining the objective of our penalized regression, we derive the parameter estimators and investigate their properties. For example, we show that the variance-covariance matrix of the penalized estimators gets smaller when the tuning parameter gets larger, and the asymptotic distribution exists for both the logistic regression and the LAD regression. We also showed that after a suitable data argumentation, the penalized regression becomes the LASSO penalized regression.

Chapter 6

Conclusion

6.1 Summary of achievements

In this thesis, we introduced two new types of penalization for regression analysis. We incorporate the sample correlation coefficients into the penalty function and call them correlation adjusted penalty: CAR and CAEN.

We extend several existing results in the literature to our new penalty function. Our main results are as follows:

(i) We proposed the new form of correlation adjusted regression (CAR) and correlation adjusted elastic net (CAEN).

(ii) We derived the penalized least squares and the penalized MLE of the regression parameters for both CAR and CAEN.

(iii) We showed that when the tuning parameter gets larger, the variance-covariance matrix of the estimator gets smaller. Therefore the mean squared error of the estimator is likely smaller than the mean squared error of the ordinary least squares estimator. This improves the performance of parameter estimation.

(iv) We showed that after using suitable data argumentation, the CAR regression is equivalent to the ordinary least squares regression, and the CAEN regression is equivalent to the LASSO regression. Therefore many properties and calculations of CAR and CAEN could be derived after data argumentation from the ordinary least squares regression and LASSO regression. Both the OLS regression and LASSO regression are well studied in the literature.

(v) We examined both CAR and CAEN for different types of regression analysis: the least squares regression for continuous responses, the logistic regression for binary responses, and the least absolute deviation (LAD) regression for continuous responses. The LAD regression is thought to be more robust than the OLS regression.

(vi) We derived the asymptotic properties of the penalized estimators for both the logistic regression and the LAD regression.

6.2 Future research

Penalized regression is very important and many different forms of penalty functions are being introduced. Penalized regression has widely spread applications in many fields, including genetics and other medical studies.

Theoretical results obtained in Chapter 4 and Chapter 5 have been submitted for publication, see Tan and Wang (2012a) and Tan and Wang (2012b). We believe there are many other approaches about penalized regression, so we would try to continue our future research in this direction. One major project we plan to work on is to compare our new penalization methods with other types of penalized regression methods by means of simulation and real data sets.

Bibliography

Alheety, M. and Ramanathan, T. (2009). Confidence interval for shrinkage parameters in ridge regression. *Communications in Statistics - Theory and Methods*, **38** 3489–3497.

Anbari, M. and Mkhadri, A. (2008). Penalized regression combining the L_1 norm and a correlation based penalty. *Research Report, Institue National de Recherche en Informatique et en Automatique*, **6746** 1–32.

Barker, L. and Brown, C. (2001). Logistic regression when binary predictor variables are highly correlated. *Statistics in Medicine*, **20** 1431–1442.

Bondell, H. and Reich, B. (2008). Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with oscar. *Biometrics*, **64** 115–123.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, **96** 1348–1360.

Frank, I. and Friedman, J. (1993). A statistical view of some chemometrics regression tools. *Technometrics*, **35** 109–135.

Gao, S. and Shen, J. (2007). Asymptotic properties of a double penalized maximum likelihood estimator in logistic regression. *Statistics & Probability Letters*, **77** 925–930.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, **12** 55–67.

Knight, K. and Fu, W. (2000). Asymptotics for lasso-type estimatiors. *The Annals of Statistics*, **28** 1356–1378.

Kutner, M. H., Nachtsheim, C. J., Neter, J. and Li, W. (2005). *Applied linear statistical models (Fifth edition)*. McGraw-Hill/Irwin, New York.

Kwon, S. and Kim, Y. (2012). Large sample properties of the scad-penalized maximum likelihood estimation on high dimensions. *Statistica Sinica*, **22** 629–653.

Kyung, M., Gill, J., Ghosh, M. and Casella, G. (2010). Penalized regression, standard errors, and bayesian lassos. *Bayesian Analysis*, **5** 369–412.

Le Cessie, S. and Van Houwelingen, J. C. (1992). Ridge estimators in logistic regression. *Applied Statistics*, **41** 191–201.

Li, Q. and Lin, N. (2010). The bayesian elastic net. *Bayesian Analysis*, **5** 151–170.

Mansson, K. and Shukur, G. (2011). On ridge parameters in logistic regression. *Communications in Statistics - Theory and Methods*, **40** 3366–3381.

Park, T. and Casella, G. (2008). The bayesian lasso. *Journal of the American Statistical Association*, **103** 681–686.

Ryan, T. (2009). *Modern regression methods (Second edition).* John Wiley & Sons, Hoboken, NJ.

Tan, Q. and Wang, X. (2012a). Correlation adjusted elastic net for regression analysis. *(Submitted).*

Tan, Q. and Wang, X. (2012b). Correlation adjusted penalization for regression analysis. *(Submitted).*

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of Royal Statistical Society Series B*, **58** 267–288.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J. and Knight, K. (2005). Sparsity and smoothness via fused lasso. *Journal of Royal Statistical Society Series B*, **67** 91–108.

Tutz, G. and Ulbricht, J. (2009). Penalized regression with correlation-based penalty. *Statistics and Computing*, **19** 239–253.

Ulbricht, J. and Tutz, G. (2008). Boosting correlation based penalization in generalized linear models. *Recent Advances in Linear Models and Related Areas Essays in Honour of Helge Toutenburg* 165–180.

Xu, J. and Ying, Z. (2010). Simultaneous estimation and variable selection in median regression using lasso-type penalty. *Annals of the Institute of Statistical Mathematics*, **62** 487–514.

Xu, Z., Zhang, H., Wang, Y., Chang, X. and Liang, Y. (2010). $L_{1/2}$ regression. *Science China - Information Sciences*, **53** 1159–1169.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, **101** 1418–1429.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B*, **67** 301–320.

Zou, H. and Zhang, H. (2009). On the adaptive elastic-net with a diverging number of parameters. *The Annals of Statistics*, **37** 1733–1751.

Appendix A

List of symbols

- $X_1, X_2, \cdots, X_p$: predictor variables
- Y: response variable
- ϵ: random error
- $\gamma_{i,j}$: correlation coefficient between X_i and X_j
- A^T: transpose matrix of matrix A
- A^{-1}: inverse matrix of the matrix A
- $\mathbf{X}$: design matrix
- $\mathbf{Y}$: vector of random responses
- H: hat matrix
- $L(\beta_0, \beta_1, \cdots, \beta_p)$: likelihood function

- $\ell(\beta_0, \beta_1, \cdots, \beta_p)$: log-likelihood function

- V: column vector, with i^{th} element being 1 if $\beta_i > 0$ and -1 if $\beta_i < 0$

- I: identity matrix

- $Var(\mathbf{Y})$: variance-covariance matrix of the random vector $\mathbf{Y}$

- $I(\beta_0)$: Fisher information evaluated at β_0

- $S(\beta)$: score function

- $\xrightarrow{P}$: convergence in probability

- $\xrightarrow{d}$: convergence in distribution

Appendix B

List of terms

- PRESS: Prediction Sum of Squares
- CAR: Correlation Adjusted Regression
- CAEN: Correlation Adjusted Elastic Net
- LAD: Least Absolute Deviation
- VIF: Variance Inflation Factor
- LINE: Linearity, Independence, Normality, Equal variance
- MLE: Maximum Likelihood Estimator
- OLS: Ordinary Least Squares
- MSE: Mean Squared Error
- BLUE: Best, Linear, Unbiased Estimator

- PLS: Penalized Least Squares
- LASSO: Least Absolute Shrinkage and Selection Operator
- SCAD: Smoothly Clipped Absolute Deviation
- OSCAR: Octagonal Shrinkage and Clustering Algorithm for Regression
- bias: bias of an estimator (expected estimator minus its parameter)
- trace: trace of a matrix (sum of diagonal elements)
- BIC: Bayesian Information Criterion
- AIC: Alkaike Information Criterion
- CV: Cross Validation
- CJRR: Canadian Joint Replacement Registry