Improved Ridge M-Estimators

Mohammad Arashi
University of Shahrood

Mina Norouzirad*
University of Shahrood

Abstract

The focus of this approach is on parameter estimation in multiple regression model in the presence of multicollinearity and outliers. Some improved ridge M-estimators are define and their performance is evaluated in a real example.

Keywords: M-Estimator; Multicollinearity; Outliers; Ridge regression; Shrinkage M-estimator.
Mathematics Subject Classification [2010]: 62G08 , 62J07

1 Introduction

A traditional linear regression model has form

$$\boldsymbol{y}_n = (y_1, \ldots, y_n)^T = X_n\boldsymbol{\beta} + \boldsymbol{\epsilon}_n, \quad \boldsymbol{\epsilon}_n = (\epsilon_1, \ldots, \epsilon_n)^T, \tag{1}$$

where $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)^T$ is the vector of unknown (regression) parameters, X_n is an $n \times p$ (design) matrix of known regression constants, $n > p \geq 1$, and the ϵ_is are errors.

There are four assumptions that must be verified before implementing the model: (i) linearity and additivity of the relationship between dependent and independent variables, (ii) statistical independence of the errors, (iii) homoscedasticity, and (iv) normality of the error distribution. When all of the assumptions are true, the best estimator for unknown parameter $\boldsymbol{\beta}$ is the ordinary least squares (OLS) estimator defined as $\hat{\boldsymbol{\beta}}_n^{OLS} = (X_n^T X_n)^{-1} X_n^T \boldsymbol{y}_n$

In real world, we may encounter a data set that doesn't satisfy one or more of the above assumptions, resulting on inappropriateness of the OLS method. Sometimes, there exist highly correlated two or more variables in collection of predictors in a regression setup. This phenomena is called multicollinearity that has been studied by many researchers in different aspects. Horel and Kennard [1] introduced the ridge regression approach to combat multicollinearity, which was already known as Tikhonov regularization. Another common problem in regression analysis is to take normality assumption for the errors, when they are not so in practice, like as fat tailed distributions, that can produce outliers. When outliers exist in the data, the use of robust estimators reduces their effects. When the regressors are fixed, so only allowing for outliers in the dependent variable (the response), it is suggested to use M-estimation, which introduced by Huber [2].

*Speaker