Journal of Multivariate Analysis 123 (2014) 68–84

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

A ridge regression estimation approach to the measurement error model[☆]

A.K.Md. Ehsanes Saleh[a], Shalabh[b,*]

[a] Carleton University, Ottawa, Canada

[b] Department of Mathematics & Statistics, Indian Institute of Technology, Kanpur - 208 016, India

ARTICLE INFO

Article history:
Received 24 September 2012
Available online 26 August 2013

AMS subject classifications:
62J05
62J07

Keywords:
Linear regression model
Measurement error
Multicollinearity
Reliability matrix
Ridge regression estimators
Shrinkage estimation
Stein type estimators
Preliminary test estimator

ABSTRACT

This paper considers the estimation of the parameters of measurement error models where the estimated covariance matrix of the regression parameters is ill conditioned. We consider the Hoerl and Kennard type (1970) ridge regression (RR) modifications of the five quasi-empirical Bayes estimators of the regression parameters of a measurement error model when it is suspected that the parameters may belong to a linear subspace. The modifications are based on the estimated covariance matrix of the estimators of regression parameters. The estimators are compared and the dominance conditions as well as the regions of optimality of the proposed estimators are determined based on quadratic risks.

1. Introduction

The standard assumption in the linear regression analysis is that all the explanatory variables are linearly independent. When this assumption is violated, the problem of multicollinearity enters into the data and it inflates the variance of an ordinary least squares estimator of the regression coefficient, see [28] for more details. Obtaining the estimators for multicollinear data is an important problem in the literature. The ridge regression estimation due to Hoerl and Kennard [13] works well in multicollinear data. The ridge estimators under the normally distributed random errors in a regression model have been studied by e.g., [31,18,19,6,12,3] etc. The details of development of other approaches and the literature related to the ridge regressions are not within the scope of this paper.

Another fundamental assumption in all statistical analyses is that all the observations are correctly observed. When this assumption is violated, the measurement errors creep into the data. Then the usual statistical tools tend to loose their validity, see [8,7] for more details. An important issue in the area of measurement errors is to find the consistent estimators of the parameters which can be accomplished by utilizing some additional information from outside the sample. In the context of multiple linear regression models, the use of additional information in the form of a known covariance matrix of

[☆] The detailed version of this paper is available on the webpage of the second author at http://home.iitk.ac.in/~shalab/research.htm. The authors are grateful to the referees and the associate editor for their comments.

* Corresponding author.
E-mail addresses: esaleh@math.carleton.ca (A.K.Md.E. Saleh), shalab@iitk.ac.in, shalabh1@yahoo.com (Shalabh).

0047-259X/$ – see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jmva.2013.08.014

measurement errors and a known matrix of reliability ratios, both associated with explanatory variables, has been studied, see e.g., [9,21–26,30,34,16,37] etc.

When the problem of multicollinearity is present in the measurement error ridden data, then an important issue is how to obtain the consistent estimators of regression coefficients. One simple idea is to use the ridge regression estimation over the measurement error ridden data. An obvious question that crops up is what happens then? In this paper, we attempt to answer such questions.

It is well known that Stein [38,14] initially proposed the Stein estimator and positive-rule estimators. The preliminary test estimators were proposed by Bancroft [4]. On the other hand, ridge regression estimators were proposed by Hoerl and Kennard [13] and they combat the problem of multicollinearity for the estimation of regression parameters. Saleh [29, Chapter 4] proposed "quasi-empirical Bayes estimators". So we have considered five quasi-empirical Bayes estimators by weighing the unrestricted, restricted, preliminary test and Stein-type estimators by the ridge "weight function". The resulting estimators are studied in measurement error models. The quadratic risks of these estimators have been obtained and optimal regions of superiority of the estimators are determined.

The plan of the paper is as follows. We describe the model set up in Section 2. The details and development of the estimators are presented in Section 3. The comparison of estimators over each other is studied and their dominance conditions are reported in Section 4. The summary and conclusions are placed in Section 5 followed by the references.

2. The model description

Consider the multiple regression model with measurement errors

$$Y_t = \beta_0 + \boldsymbol{x}_t'\boldsymbol{\beta} + e_t, \qquad \boldsymbol{X}_t = \boldsymbol{x}_t + \boldsymbol{u}_t, \quad t = 1, 2, \ldots, n \tag{2.1}$$

where β_0 is the intercept term and $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)'$ is the $p \times 1$ vector of regression coefficients, $\boldsymbol{x}_t = (x_{1t}, x_{2t}, \ldots, x_{pt})'$ is the $p \times 1$ vector of set tth observations on true but unobservable p explanatory variables that are observed as $\boldsymbol{X}_t = (X_{1t}, X_{2t}, \ldots, X_{pt})'$ with $p \times 1$ measurement error vector $\boldsymbol{u}_t = (u_{1t}, u_{2t}, \ldots, u_{pt})'$, u_{it} being the measurement error in the ith explanatory variable x_{it} and e_t is the response error in the observed response variable Y_t. We assume that

$$(\boldsymbol{x}_t', e_t, \boldsymbol{u}_t) \sim N_{2p+1}\left\{(\boldsymbol{\mu}_x', 0, \boldsymbol{0}')', \text{BlockDiag}(\Sigma_{xx}, \sigma_{ee}, \Sigma_{uu})\right\} \tag{2.2}$$

with $\boldsymbol{\mu}_x = (\mu_{x_1}, \mu_{x_2}, \ldots, \mu_{x_p})'$, σ_{ee} is the variance of e_t's whereas Σ_{xx} and Σ_{uu} are the covariance matrices of x_t's and u_t's respectively. Clearly, $(Y_t, \boldsymbol{X}_t')'$ follows a $(p+1)$-variate normal distribution with mean vector $(\beta_0+\boldsymbol{\beta}'\boldsymbol{\mu}_{\mathbf{x}}, {\boldsymbol{\mu}_{\mathbf{x}}}')'$ and covariance matrix

$$\begin{pmatrix} \sigma_{ee} + \boldsymbol{\beta}'\Sigma_{xx}\boldsymbol{\beta} & \boldsymbol{\beta}'\Sigma_{xx} \\ \Sigma_{xx}\boldsymbol{\beta} & \Sigma_{xx} + \Sigma_{uu} \end{pmatrix}. \tag{2.3}$$

Then the conditional expectation of Y_t given $\boldsymbol{X}_t$ is

$$E(Y_t|X_t) = \gamma_0 + \boldsymbol{\gamma}'\boldsymbol{X}_t \tag{2.4}$$

where $\gamma_0 = \beta_0 + \boldsymbol{\beta}'(I_p - K_{xx}')\boldsymbol{\mu}_x$, $\boldsymbol{\gamma} = K_{xx}\boldsymbol{\beta}$, $\boldsymbol{\beta} = K_{xx}^{-1}\boldsymbol{\gamma}$, and $K_{xx} = \Sigma_{XX}^{-1}\Sigma_{xx} = (\Sigma_{xx} + \Sigma_{uu})^{-1}\Sigma_{xx}$ is the $p \times p$ matrix of reliability ratios of X, see [9].

Our basic problem is the estimation of $\boldsymbol{\beta}$ under various situations beginning with the primary estimation of $\boldsymbol{\beta}$ assuming Σ_{uu} is known.

Let

$$S = \begin{pmatrix} S_{YY} & S_{YX} \\ S_{XY} & S_{XX} \end{pmatrix} \tag{2.5}$$

where

(i) $S_{YY} = (\boldsymbol{Y} - \bar{Y}\mathbf{1}_p)'(\boldsymbol{Y} - \bar{Y}\mathbf{1}_p)$, $Y = (Y_1, Y_2, \ldots, Y_n)'$, $\mathbf{1}_n = (1, 1, \ldots, 1)'$.
(ii) $S_{XX} = ((S_{X_iX_i}))$, $S_{X_iX_i} = (\boldsymbol{x}_i - \bar{X}_i\mathbf{1}_n)'(\boldsymbol{x}_i - \bar{X}_i\mathbf{1}_n)$
(iii) $S_{\mathbf{X}_iY} = (\boldsymbol{X}_i - \bar{X}_i\mathbf{1}_n)'(\boldsymbol{Y}_i - \bar{Y}_i\mathbf{1}_n)$, $S_{XY} = (S_{X_1Y}, S_{X_2Y}, \ldots, S_{X_pY})'$
(iv) $\bar{X}_i = \frac{1}{n}\sum_{t=1}^n X_{it}$, $\bar{Y} = \frac{1}{n}\sum_{t=1}^n Y_t$.

Gleser [9] showed that the maximum likelihood estimators of γ_0, $\boldsymbol{\gamma}$ and σ_{zz} are just the naive least squares estimators, viz.,

$$\tilde{\gamma}_{0n} = \bar{Y} - \tilde{\boldsymbol{\gamma}}_n'\bar{\mathbf{X}}, \qquad \tilde{\boldsymbol{\gamma}}_n = S_{XX}^{-1}S_{XY} \quad \text{and} \quad \tilde{\sigma}_{zz} = \frac{1}{n}(Y - \tilde{\gamma}_{0n}\mathbf{1}_n - \tilde{\boldsymbol{\gamma}}_n'\boldsymbol{X})'(Y - \tilde{\gamma}_{0n}\mathbf{1}_n - \tilde{\boldsymbol{\gamma}}_n'\boldsymbol{X}) \tag{2.6}$$

provided

$$\tilde{\sigma}_{ee} = \tilde{\sigma}_{zz} - \tilde{\boldsymbol{\gamma}}_n'K_{xx}^{-1}\Sigma_{uu}\tilde{\boldsymbol{\gamma}}_n \geq 0. \tag{2.7}$$