Empirical Bayes and Bayes for Ridge regression

Muhannad Faiz Al-Sadoun / Department of statistics / Al-qadisiya universit

Abstract

In this paper , we consider the problem of estimating the regression parameters in a multiple linear regression model , when the multicollinearity is present under the assumption of normality , we present two types , empirical Bayes and Bayes . One of them shrinks the least squares (LS) estimator towards the principal component .The second type is a hierarchical Bayesian model . A simulation example is given .

1- Introduction

Shrinkage estimation in multiple regression has been interested in a topic of since ridge regression was introduced by Horel and kennard (1970) [4,2] . The first suggestion by James and stein (1961) , proposed the so-called ridge regression method which is unaffected by the multicollinearity among the many independent variables [7,6] . There are various another for ridge regression , e.g Sclove (1968) and Baranckik (1973) , have justified Shrunken estimators over least squares on grounds of reduced mean squared error [7] . A more general method called continuum regression has been proposed by Stone and Brooks (1990) [7] . This procedure depend on a parameter , say γ which is recommended to be cross validation [7] . However , except for two special values of γ , (0 and 1) . Sundberg (1993) [8] and Bjorkstrom and Sundberg (1996) [1] have shown that it is equivalent to ridge regression . The Bayesian Theorem has been employed successfully by Lindley and Smith (1972) [6] , Novick etal , (1972) [6] , Zellner (1971) [10] , Box and Tiao (1973) [9] , and Goldstain and Smith (1974) [9] , to name a few . In most of these treatments , shrinkage is towards zero . Tatsuya and Srivastave [9] has been improved empirical Bayes for ridge regression estimation by three methods .

This article deals with Bayes and empirical Bayes for estimating ridge regression coefficients in a normal multiple linear regression model under multicollinearity by the simulation for Normality data .

The ridge regression estimator and multicollinearity in section 2 with normal distribution for error term . In section 3 , the Bayes case is

developed with exchangeable normal prior on the ridge regression coefficients in section 4 , the empirical Bayes treatment is develop in section 5 , contain some results of the simulation and the final section represents a general discussion .

2- Ridge regression

The ordinary least squares (OLS) method faced problems in estimation accuring when there's a great correlation between independent variables which called Multicollinearity problem . Horel & Kennard (1970) [4] suggest a method to treat these problems which called Ridge regression (R.R) , which first it subtract the dependant variable and independent variables from the arithmetic mean for each variable , so Ridge regression estimators will be as following :-

$$\hat{\beta}_R^{*} = (x^{*/} x^{*} + kI)^{-1} x^{*/} y \qquad \text{.....(2.1)}$$

Where (kI) represents Multiply constant value (K) by unit matrix This method briefed of addition of the constant value (k) to the diagonal elements of the matrix ($x^{*'} x^{*}$) and the method named ordinary Ridge regression (ORR) , Horel declared the possibility to obtain the most accurate estimations by adding different values to the diagonal elements of the matrix ($x^{*'} x^{*}$) and this method called Generalized Ridge regression (GRR). Then $\hat{\beta}_R^{*}$ estimators will be as following :-

$$\hat{\beta}_R^{*} = (x^{*/} x^{*} + k)^{-1} x^{*/} y \qquad \text{.....(2.2)}$$

Such that

$$k = \begin{pmatrix} k_1 & 0 \ldots\ldots 0 \\ 0 & k_2 \ldots\ldots 0 \\ 0 \ldots 0 \ldots\ldots & k_t \end{pmatrix} \quad , \quad k_1 \neq k_2 \neq \ldots\ldots \neq k_t$$

And some of the method properties are :-
1- Ridge regression estimators will be biased as following :-

$$\hat{\beta}_R^{*} = (x^{*/} x^{*} + kI)^{-1} x^{*/} y$$
$$= [(x^{*/} x^{*} + kI)^{-1} x^{*/} (x^{*} \underline{\beta} + \underline{u})]$$
$$= [(x^{*/} x^{*} + kI)^{-1} x^{*/} x^{*} \underline{\beta} + (x^{*/} x^{*} + kI)^{-1} x^{*/} \underline{u}]$$

$\because$ $E \underline{u} = 0$

$$E (\underline{\beta}^{\wedge *}_R) = (x^{*/} x^{*} + kI)^{-1} x^{*/} x^{*} \underline{\beta} \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots (2.3)$$

Let $Z = (x^{*/} x^{*} + kI)^{-1} x^{*/} x^{*}$

$E (\underline{\beta}^{\wedge *}_R) = Z \underline{\beta}$.

And through above , we notice that biased value is Z .

2- The variance of Ridge regression estimators calculated from the following :

$$var(\underline{\beta}^{\wedge *}_R) = var [(x^{*/} x^{*} + kI)^{-1} x^{*/} y]$$
$$= (x^{*/} x^{*} + kI)^{-1} x^{*/} var(y) x^{*/} (x^{*/} x^{*} + kI)^{-1} \qquad \ldots\ldots\ldots\ldots\ldots (2.4)$$

3- The summation of squared error by Ridge regression (RR) is less than (OLS) method and as following :-

$$MS_e(\beta^{\wedge *}_R) = tr [var(\underline{\beta}^{\wedge *}_R)] + [Bias (\underline{\beta}^{\wedge *}_R)]^2$$
$$= \sigma^2 tr[(x^{*/} x^{*} + kI)^{-1} x^{*/} x^{*} (x^{*/} x^{*} + kI)^{-1}] + \underline{\beta}'(Z-I)(Z-I)\underline{\beta} \quad \ldots.(2.5)$$

and from above it is clear that (K) value is an increasing function in (K) and about the variance , it is a decreasing function in (K), so we accept with one limited value of biased , and when biased value increased , variance value will be decreased , so it will be less than the summation of the squared errors to (OLS) method .When concerning (k) value ,which will be added to the diagonal elements of the matrix ($X^{*/}X^{*}$) , must be etermined We should depend through this paper on Horel method , which said that optimum value of (k) calculated according the following formula , when using (ORR) method :

$$K = \frac{k\ s^2_e}{\underline{\beta}^{\wedge/}_{Ls}\ \underline{\beta}^{\wedge}_{Ls}} \qquad \text{……..……..……(2.6)}$$

And when applying (GRR) method it is possible to find many optimum values for (k) according to the following :

$$K_i = \frac{S^2_e}{b^2_i} \qquad i = 1 , 2 , \ldots\ldots, k \qquad \text{……...(2.7)}$$

3- Bayes Ridge Regression

If the model in standard form as :

$$\underline{y} = \underline{X}\ \underline{ß} + \underline{u} \qquad \text{……..(3.1)}$$

Such that:

$\underline{y}$:- be a vector containing n independent observations $y_1, y_2, \ldots\ldots y_n$ on the response variable .

X :- $n_{*}p$ matrix of observations on (P) predictor and multicollinearity between the variables .

$\underline{ß}$: - $(ß_1 , ß_2 , \ldots.., ß_p)^T$ be the vector of the regression coefficients .

$\underline{u}$:- is the n_{*1} error vector , with each $u \sim N(0,\sigma^2)$.

Note that in this model it is assumed that the data have translated to local the origin at the grand mean , so that there will be no intercept among the regression coefficients .
There two stages for estimating Bayes Ridge regression :

The first stage :-

The hierarchical model , we have assumed a priori that these are independently and identically normally distribute thus ,

$$ß / \mu , T^2 \sim N(\mu 1 , T^2 I_p) \quad \ldots\ldots(3.2)$$

Where: $1 = (1,1,\ldots\ldots,1)^T$ and I_P is the identity matrix of order (p) .
To complete the hierarchical Bayesian model . we need for determining the prior distributions for μ and T^2 .At this level , μ and T^2 are assumed to be independent a priori and have priors , respectively , $p(\mu)$ and $p(T^2)$. The matrix (X) , the value σ^2 , and the forms of $p(\mu)$ and $p(T^2)$ are assumed known .

The second stage

The posterior density of β conditional on μ and T^2 is multivariate normal $N(\beta^*,\Omega^*_\beta)$ with

$$\beta^* = [I_p - (x^/ x)^{-1} \varphi] b + (x^/ x)^{-1} \varphi 1 \mu \quad \ldots\ldots(3.3)$$

$$\Omega^*_\beta = \sigma^2 [(x^/ x)^{-1} - (x^/ x)^{-1} \varphi (x^/ x)^{-1}] \quad \ldots\ldots(3.4)$$

$$\text{Where } \varphi = [(x^/ x)^{-1} + (T^2/\sigma^2) I_p]^{-1} \quad \ldots\ldots(3.5)$$

b is the ordinary Ridge regression (ORR) or Generalized Ridge regression (GRR) if β is of best variance value .

4- Empirical Bayes Ridge regression [9]

Suppose that , the prior distribution

$$\beta \sim N_p (H^t_2 \alpha , \sigma^2 \lambda I_p) \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (4.1)$$

Such that :

λ is unknown

H_2 is an orthogonal matrix

Then posterior distribution

$$\beta / \hat{\beta} \sim N_p [\hat{\beta} (\lambda , \alpha) , \sigma^2 (x^{/} x + \lambda^{-1} I)^{-1}] \qquad \ldots\ldots\ldots\ldots\ldots (4.2)$$

Where $\hat{\beta}^{\beta} (\lambda , \alpha)$ is the Bayes estimator of β given by :-

$$\hat{\beta}^{\beta} (\lambda , \alpha) = (x^{/} x + \lambda^{-1} I)^{-1} x^{/} x (\hat{\beta} - H^{/}_2 \alpha) + H^{/}_2 \alpha$$

$$= \hat{\beta} - (I + `\lambda x^{/} x)^{-1} (\hat{\beta} - H^{/}_2 \alpha) \qquad \ldots\ldots\ldots\ldots\ldots (4.3)$$

Since α and λ are unknown , they need to be estimated .

First , α may be estimated by the weighted least squares estimator :

$$\hat{\alpha} = (H_2 x^{/} x H^{/}_2) H_2 x^{/} x \hat{\beta} \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots (4.4)$$

Which can be obtained by minimizing the weighted least squares loss $(\hat{\beta} - H^{/}_2 \alpha) x^{/} x (\hat{\beta} - H^{/}_2 \alpha)$. We see that $\hat{\alpha} = H_2 \beta$.

Since the principle component (PC) regression estimator $\hat{\beta}^{pc}$ of β is given by : $\hat{\beta}^{pc} = H^{/}_2 H_2 \hat{\beta}$ $\qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (4.5)$

We observe that $H^{/}_2 \alpha = \hat{\beta}^{pc}$ substituting $H^{/}_2 \alpha = \hat{\beta}^{pc}$ into $\hat{\beta}^{\beta} (\lambda , \alpha)$, we get the estimator :

$$\hat{\beta}^{\beta} (\lambda , \hat{\alpha}) = \hat{\beta} - (I + `\lambda x^{/} x)^{-1} (\hat{\beta} - \hat{\beta}^{pc}) \qquad \ldots\ldots\ldots\ldots\ldots\ldots (4.6)$$

A reasonable method to estimate λ is from the marginal distribution of $\hat{\beta}$ - using the sample moments , we propose an estimator which we call an empirical Bayes estimator . Let λ^{*} be a root of the equation

$(\hat{\beta} - \hat{\beta}^{pc})' \{ (x' x)^{-1} + \lambda I \}^{-1} (\hat{\beta} - \hat{\beta}^{pc}) = (p- q-2) * s / (n+2)$ ...(4.7)
and λ_0 is the root of the equation :

$$\sum_{i=1}^{p-q} \frac{d_i - d_{p-q}}{d_i + \lambda_0} = (p-q-2)/2 \quad \text{.....(4.8)}$$

Then we propose the estimator $\hat{\lambda}_{EB}$ of λ , given by the maximum of λ^* and λ_0 , that is ,

$\hat{\lambda}_{EB} = \max (\lambda^* , \lambda)$ ……...……(4.9)
substituting $\hat{\alpha}$ and $\hat{\lambda}_{EB}$ into (4.3) , we get the empirical Bayes ridge regression estimator EB

$\hat{\beta}^{EB} = \hat{\beta}(\hat{\lambda}_{EB} , \hat{\alpha}) = \hat{\beta} - (I+ \hat{\lambda}_{EB} x' x)^{-1} (\hat{\beta} - \hat{\beta}^{pc})$ ……..….(4.10)

Which shrinks the (LS) estimator $\hat{\beta}$ towards to pc estimator $\hat{\beta}^{pc}$. It is known that the principle component estimator and the ridge regression estimation are useful in predicting a response variable in the presence of multicollinearity . It is interesting to note that both methods of ridge regression and principle components are incorporated in the proposed estimator $\hat{\beta}(\hat{\lambda}_{EB} , \hat{\alpha})$.

5-Simulation study :

Simulation is a numerical technique for conducting experiments on a digital computer , which involve certain types of mathematical and logical models that describe the system behavior .
It is often viewed as a "Method of lost resort " to be used when everything else has failed , software building and technical developments have made simulation one of the most widely used and accepted tools for designer in system analysis and operational research .

In this paper (20) observations had been generated for all of Y , X_1 and X_2 with noticing that there is a relationship between the independent variables , where :

$Y \sim N(0,1)$
$X_1 \sim N(0,1)$
$X_2 \sim N(0,1) + 2X_1$

And the results will be as following :

5.1- Ordinary Least squares

(OLS) method had been applied without estimation the constant parameter and the results of the estimation will be as follow:

$y = -0.4957\ x1 + 0.4354x2$

s.d	0.4857154	0.23847
t	-1.021	1.826

$s^2_e = 1.1071$, $R^2 = 60.8\ \%$, $F = 0.3326$

By noticing the above results , it is clear that the variance for high parameter and for a general sample through (variance errors ,R^2,F)
It is not significant although that the significant parameter (β_1,β_2) which estimated by t value , because there is a relationship between the independent variables , which is called the multicollinearity and the researcher used Ridge regression method in two cases :

A-Ordinary Ridge regression

In order to identify on the type of the functional relationship , one value of (k) had been added to the main diagonal of the matrix $(x^{*\prime} x^{*})$ according to equation number (2 .1) , where the added value was (k = 0.5) , and according to that it was possible to determine the estimated value for the function after adding (k) value where it became in the following shape :

$$y = -0.1213\, x1 + 0.0737\, x2$$
$$s.d \quad 0.14257 \quad 0.006999$$
$$t \quad 0.8508 \quad 10.5300$$
$$s^2_e = 1.078$$

we notice from the above estimation equation that (t) values is not significant for (x1) and significant for(x_2) . and error variance value becomes $s^2 = 1.078$, and by comparing with the estimation by (OLS) method , we realize a noticeable improvement in model by using Ridge approach by increasing the significance of the parameters and decreasing the variance despite the bias value which included in this model .

B- Generalized Ridge Regression

Generalized Ridge Regression had been applied on the data of this research , where different values of k had been added to the main diagonal of the matrix

according to equation (2.2) as shown in the following $(x^{*\prime} x^{*})$ matrix :-

by using the above values of (K) it

$$k = \begin{pmatrix} 0.6 & 0 \\ 0 & 0.7 \end{pmatrix}$$

was possible to estimate the function by using (GRR) and it was as following :

$\hat{y}$ = 0.107 X1+0.0521 X2
s.d 0.2768 0.2372
t 0.3865 0.2196
s^2 = 1.0794

We notice from the estimated model above , there is convergence in the results between (GRR) and (ORR) , where we notice that the values of (s^2_e , t , s.d) are very convergent by both methods except the value of variable(X2) has been height its value ,despite the decrement of variance in both cases by (OLS) method , but depending on variance , we notice that (ORR) is better than (GRR) , because it's variance is lower than (GRR).

5.2- Bayes Ridge regression :

Bayes approach had been applied according to equations (3.1)-(3.2) and the results were at two stages :-

First stage : through this stage the prior distribution had been determined for every parameters , where it had been obtained by partitionate the observations into equal groups and through these groups the parameters and the variance of the model will be estimate in both (GRR , ORR) methods according to the following table :

Table no.(1) shows parameters and variance estimation by (GRR) and (ORR)

Observations		β_1 ORR	β_2	β_1 GRR	β_2
1-5	β_i	-0.0485	0.0621	-0.0397	0.0459
	s.d	0.23856	0.129098	0.1541	0.1418
	s^2e	1.0271		1.0273	
6-11	β_i	-0.2150	-0.2616	-0.2095	-0.2373
	s.d	0.3500	0.201	0.1747	0.1571
	s^2e	1.4481		1.4693	
11-15	β_i	-0.3468	-0.4343	-0.3343	-0.3928

	s.d	0.29215 0.22316	0.3527 0.3041
	s^2e	0.8263	0.8933
16-20	β_i	0.2827 0.1438	0.2636 2.1465
	s.d	0.32798 0.1140445	0.1998 0.1762
	s^2e	1.669	1.6889

And through the above table, best model had been selected according to minimum of maximum variance to the models which through it the prior distribution had been selected. And test had been done according minimum maximum (minimax) variance which is the observations 11-15accordingto(GRR)approach:

Observations		β_1 ORR β_2
11-15	β_i	-0.3343 -0.3928
	s.d	0.3527 0.3041
	s^2e	0.8933

and after the prior distribution , we depend the formula no. (3.3)-(3.5) to determine the posterior distribution and the results were as follows :

$Y = -0.4023X_1 - 0.0823X_2$
s.d 0.01132 0.000245 $S^2e = 0.8933$

By noticing the above results , it is shown that the parameters variance is low and in general we notice an improvement in the next model for the values and the parameters variance .

5.3-EmpiricalBayesRidgeregression

Empirical Bayes had been applied according to equation (4.10) and the results were as follows :

$y = 0.411018\ X_1 - 0.147097\ X_2$
s.d 25.3174 99.3893 , $s^2_e = 1.16518$

Through the above results , we notice an increasing in the parameters variance value and also the errors variance , compare to Bayes approach we notice that the results we obtain is better than Empirical Bayes results and through the variance and errors variance , so Bayes approach is better than Empirical Bayes , because the determination the prior distribution i.e, the prior information on the phenomena before determine the posterior distribution for the phenomena and it could not control on phenomena through sets of data in a certain moment .

6-Conclusions :

Through the empirical study it shows the following :

1- (OLS) method , displayed not significant results , because the model suffer from the problem of multicollinearity , and (GRR,ORR) had been applied and from the results depend on errors variance in both methods , (ORR) is better than (GRR) , because of (ORR) variance decrement compare to (GRR) .

2 – Bayes approach for Ridge regression (ORR) and (GRR) showed better results than the ordinary method , because it depends on prior distribution to the phenomena and then the posterior distribution , and showed that (GRR) Bayes is better than (ORR) Bayes in the stage of determining the posterior distribution .

3- Empirical Bayes didn't showed good results , because increasing the value of parameters variance and errors variance .[and by iterative empirical Bayes approach it is possible to obtain results near than Bayes approach] [3] so , (GRR) Bayes approach is the best in estimation .

References:

1- BjÖrkstrom , A.and Sundberg , R.,1996.Continuum regression is not always continuous .J.R. Statist . Soc., B , 58 , 703-710.
2- Fassil NEBEBE and T.W.F. stround. 1986 "Bayes and empirical Bayes shrinkage estimation of regressio coefficient" . The canadian Journal of Statistics , 14 , 4 , p.267-280.
3- G.bbnos , D., 1981 ," A simulation study of some Ridge Estimator " JASA , 76 , 131-139 .
4- Horel , A.E and Kennarad , R.W. (1970a). Ridge regression Biased Estimation for nonorthogonal .Technometrics , 12 , 55-66.
5- Krishna , K. 1980 . "Some finite sample properties of generalized ridge regression estimators ". The canadian Journal of statistics 8,1,47-58.
6- Lindley , D.V. and Smith , A.F.M. 1972, Bayes estimates for the linear model (with discussion) . J.Roy. Statist . Soc. , 34 , 1-41.
7- Stone . M. and Brooks , R.J. 1990. Continuum regression : cross-validated sequentially constructed prediction embracing ordinary least squares , partial least square and principle components regression . J.R. Statist . Soc. , B, 52, 237-269.
8- Sundberg , R. 1993. Continuum regression and ridge regression J.R. Statist , Soc. , B , 55 , 653-659.
9- Tatsuya Kubokawa , and M.S.Srivastave . 2003. "Improved Empirical Bayes Ridge regression Estimators under Multicollinearity" www.ustat.utstat.toronto .

10- Zellner , A.(1971) . An introduction to Bayesian Inference in Econometrics . wiely , Newyork.