Employing Ridge Regression Procedure to Remedy the Multicollinearity Problem

Hazim M. Gorgees
Bushra A. Ali
Dept. of Mathematics/College of Education for Pure Science(Ibn AL-Haitham) / University of Baghdad

Received in:11 September 2012 Accepted in:15 October 2012

Abstract

In this paper we introduce many different Methods of ridge regression to solve multicollinearity problem in linear regression model. These Methods include two types of ordinary ridge regression (ORR_1), (ORR_2) according to the choice of ridge parameter as well as generalized ridge regression (GRR). These methods were applied on a dataset suffers from a high degree of multicollinearity, then according to the criterion of mean square error (MSE) and coefficient of determination (R^2) it was found that (GRR) method performs better than the other two methods.

Keywords : Ordinary ridge regression, Generalized ridge regression, Shrinkage estimators, Singular value decomposition, Coefficient of determination.

Introduction

In this paper we deal with the classical linear regression model $y = X\beta + \varepsilon$ ….. (1) where y is (n×1) vector of response variable, X is (n×p) matrix,(n > p)of explanatory variables, β is (p×1) vector of unknown parameters and ε is (n×1) vector of unobservable random errors, where $E(\varepsilon)=0$, $var(\varepsilon)=\sigma^2 I$. Considerable attention is currently being focused on biased estimation of the regression estimators of a linear regression model .This attention is due to the inability of classical least squares to provide reasonable point estimates when the matrix of regression variables is ill-conditioned. Despite possessing the very desirable property of being minimum variance in the class of linear unbiased estimators under the usual conditions imposed on the model, the least squares estimators can, nevertheless, have extremely large variances when the data are multicollinear which is one form of ill-conditioning. Much research, therefore, on obtaining biased estimators with better overall performance than the least squares estimator is being conducted. This paper discusses the ridge regression estimators for use with multicollinear data. In contrast to least squares, these estimators allow a small amount of bias in order to achieve a major reduction in the variance. A numerical example is included to illustrate the theoretical relationships.

The Case of Multicollinearity

The problem of multicollinearity occurs when there exists a linear relationship or an approximate linear relationship among two or more explanatory variables; two types of multicollinearity may be faced in regression analysis , perfect and near multicollinearity. As an example of perfect multicollinearity assuming that the three components of a mixture are studied by including their percentages of the total p_1,p_2,p_3 obviously these variables will have the perfect linear relationship $p_1+p_2+p_3 =100$. During regression calculations, the exact linear relationship causes a division by zero which in turn causes the calculations to be aborted. When the relationship is not exact, the division by zero does not occur and the calculations are not aborted. However the division by a very small quantity still distorts the results. Hence, one of the first steps in a regression analysis is to determine if a multicoiiinearity is a problem. Multicollinearity can be thought of as a situation where two or more explanatory variables in the data set move together. As a consequence it is impossible to use this data set to decide which of the explanatory variables is producing the observed change in the response variable. Moreover, multicollinearity can create inaccurate estimates of the regression coefficients. To deal with multicollinearity we must be able to identify its source. The source impacts the analysis, the corrections and the interpretation of linear model. The sources of multicollinearity may be summarized as follows :[1]

1- Data collection. In this case the data have been collected from a narrow subspace of the explanatory variable. The multicollinearity has been created by the sampling methodology and doesn't exist in the population .Obtaining more data on an expanded range would cure this multicollinearity problem.

2- Physical constraints on the linear model or population. This source will exist no matter what sampling technique is used. Many manufacturing or service processes have constraints on explanatory variables (as to their range),either physically, politically, or legally which will create multicollinearity moreover extreme values or outliers in the X space can cause multicollinearity.

Some multicollinearity is nearly always present, but the important point is whether the multicollinearity is serious enough to cause appreciable damage to the regression analysis. Indicators of multicollinearity include a low determinant of the information matrix X′X, a very high correlation among two or more explanatory variables, very high correlation among two or more estimated coefficients a very small (near zero)eigenvalues of the correlation matrix of the explanatory variables. Moreover the Farrar-Glauber test based on Chi square statistic may be used to detect multicollinearity. Accordingly the null hypothesis to be tested is:

H_0 : X_j are orthogonal , j= 1,2,….p

Against an alternative

H_1 :X_j are not orthogonal.

The test statistic is

$$\chi^2 = -\left[(n-1) - \frac{1}{6}(2p+5)\right] \ln |D| \qquad \text{…………..(2)}$$

Where n is the number of observations, p is the number of explanatory variables, |D| is the determinant of correlation matrix. Comparing the calculated value of χ^2 with theoretical value at p(p-1)/2 degrees of freedom and specified level of significant, we reject H_0 if the calculated value is more than the theoretical value which means that the dataset suffers from a multicollinearity problem, otherwise the null hypothesis H_0 can not be rejected.

The Shrinkage Estimators

Applying the singular value decomposition we can decompose an (n×p) matrix into three matrices as follows:

$$X=H\,D^{1/2}\,G' \qquad \text{…..…………(3)}$$

Where H is an n×p semi orthogonal matrix satisfying $H'H = I_P$, $D^{1/2}$ is a (p×p) diagonal matrix of ordered singular values of X

$d_1^{1/2} \geq d_2^{1/2} \geq \ldots\ldots \geq d_p^{1/2} > 0$, G is a (p ×p) orthogonal matrix whose columns represent the eigenvectors of $X'X$.

Accordingly, the ordinary least squares estimator of the regression parameter vector β can be written as:

$$b_{OLS} = (X'X)^{-1}X'Y = GC$$

Where $C= D^{-1/2}H'Y$ is a (p×1) vector containing the uncorrelated components of b_{OLS} [2]

The generalized shrinkage estimators will be denoted by b_{SH} may be defined as: [1]

$$b_{SH} = G\Delta C = \sum_{j=1}^{p} \overrightarrow{g_j}\delta_j c_j \qquad \text{…………..(4)}$$

where $\overrightarrow{g_j}$ is the j-th column of the matrix G, δ_j is the j-th diagonal element of the shrinkage factors diagonal matrix Δ, $0 \leq \delta_j \leq 1$, j = 1,2,…,p, c_j is the j-th element of the uncorrelated component vector C.

Ordinary Ridge Regression Estimators

One of several methods that have been proposed to remedy multicollinearity problem by modifying the method of least squares to allow biased estimators of the regression coefficients, is the ridge regression method. The ridge estimator depends crucially upon an exogenous parameter, say k called the ridge parameter or the biasing parameter of the estimator. For any k ≥0, the corresponding ridge estimator denoted by b_{RR} is defined as:

$$b_{RR} = (X'X + kI)^{-1}X'Y \qquad \text{………..(5)}$$

Where k ≥ 0 is a constant chosen by the statistician on the basis of some intuitively plausible criteria put forward by Hoerl and Kennard.[3]

It can be shown that the ridge regression estimator given in (5) is a member of the class of shrinkage estimators as follows:[2]

By using Matrix algebra and singular value decomposition approach of matrix X we get:

$$b_{RR} = (X'X + kI)^{-1} X'Y = [G(D+kI)G']^{-1}GD^{1/2}\,H'Y$$

$= G(D+kI)^{-1}G'GD^{1/2}H'Y$
$= G(D+kI)^{-1}D^{1/2}H'Y$
$= G[(D+kI)^{-1}D]D^{-1/2}H'Y$
$= G\Delta C$ .... (6)

Where $\Delta=(D+kI)^{-1}D$. Equivalently, the shrinkage factors δ_j, $j=1,\ldots,p$ of the ridge estimator have the form

$$\delta_j = \frac{d_j}{d_j+k}, \quad j=1,2,\ldots,p \qquad \text{............ (7)}$$

where d_j is the j^{th} element (eigenvalues) of the diagonal matrix D, and k is the ridge parameter.

The Generalized Ridge Regression (GRR)

In this section we suggest using the singular value decomposition technique in order to derive the generalized ridge regression estimator for the first time (as far as we know). Let G be a $(p\times p)$ orthogonal matrix with columns as eigenvectors $(g_1, g_2 \ldots g_p)$ of $X'X$ Hence, $G'(X'X)G = G'(GDG')G = D = \mathrm{diag}(d_1, d_2, \ldots d_p)$. Then we can rewrite the linear model as:

$Y = X\beta + \varepsilon$
$=(HD^{1/2})G'\beta + \varepsilon = X^*\alpha + \varepsilon$ .... (8)

Where $X^* = HD^{1/2}$, $\alpha = G'\beta$

This model is called canonical linear model or uncorrelated components model. The OLS estimate for α is given as :

$\alpha_{OLS} = (X^{*\prime}X^*)^{-1}X^{*\prime}Y = (D^{½}H'H D^{½})^{-1}X^{*\prime}Y$
$= D^{-1}X^{*\prime}Y$ ....(9)

and $\mathrm{var}(\alpha_{OLS}) = \sigma^2(X^{*\prime}X^*)^{-1} = \sigma^2 D^{-1}$.

Which is diagonal. This shows the important property of this parameterization since the elements of α_{OLS} namely $(\alpha_1, \alpha_2 \ldots \alpha_p)_{OLS}$ are uncorrelated. The ridge estimator for α is given by:

$\alpha_{RR} = (X^{*\prime}X^* + K)^{-1}X^{*\prime}Y = (D+K)^{-1}X^{*\prime}Y$ ....(10)
$= (D+K)^{-1}X^{*\prime}X^*\alpha_{OLS}$
$= (I+KD^{-1})^{-1}\alpha_{OLS}$
$= w_K\alpha_{OLS} = \mathrm{diag}\left(\frac{d_i}{d_i+k_i}\right)\alpha_{OLS}$.

Where K is a diagonal matrix with entries $(k_1, k_2 \ldots k_p)$. This estimate is known as generalized ridge estimate. The mean square error of α_{RR} is given by:

$MSE(\alpha_{RR}) = \mathrm{var}(\alpha_{RR}) + (\mathrm{bias}\ \alpha_{RR})(\mathrm{bias}\ \alpha_{RR})'$
$= \sigma^2 \mathrm{tr}(w_k D^{-1}w'_k) + (w_k - I)\alpha_{OLS}\alpha'_{OLS}(w_k - I)'$

$$= \sigma^2 \sum \frac{d_i}{(d_i+k_i)^2} + \sum \frac{k_i^2\alpha^2_{(OLS)i}}{(d_i+k_i)^2} \qquad \text{............(11)}$$

To obtain the value of k_i that minimize MSE (α_{RR}) we differentiate equation(11)with respect to k_i and equating the resultant derivative to zero thus

$$\frac{\partial MSE(\alpha_{(RR)})}{\partial k_i} = -\sigma^2 \sum \frac{d_i}{(d_i+k_i)^3} + \sum \frac{d_i k_i \alpha^2_{(OLS)i}}{(d_i+k_i)^3} = 0$$

Solving for k_i we obtain : $k_i = \dfrac{\sigma^2}{\alpha^2_{(OLS)i}}$

Since the value of σ^2 is usually unknown we use the estimate value $\sigma^{\wedge 2}$. Accordingly, when matrix K satisfies $\hat{k}_i = \dfrac{\hat{\sigma}^2}{\alpha^2_{(OLS)i}}$

$$= \text{diag}\left(\frac{\hat{\sigma}^2}{\alpha^2_{(OLS)1}}, \frac{\hat{\sigma}^2}{\alpha^2_{(OLS)2}}, \ldots \frac{\hat{\sigma}^2}{\alpha^2_{(OLS)p}}\right)$$

Then the mean square error of generalized ridge regression estimate α_{RR} attains the minimum value. The original form of ridge regression estimator can be converted back from the canonical form by:

$$b_{(GRR)} = G\,\alpha_{(RR)} \qquad \ldots\ldots\ldots\ldots\ldots(12)$$

All the basic results concerning the ordinary ridge regression estimator can be shown to hold for this more general formulation.

Choice of Ridge Parameter

The ridge regression estimator does not provide a unique solution to the problem of multicollinearity, but provide a family of solutions .These solutions depend on the value of k (the ridge biasing parameter). No explicit optimum value can be found for k. Yet, several stochastic choices have been proposed for this shrinkage parameter .Some of these choices may be summarized as follows:[4]

Hoerl and Kennard (1970), suggested a graphical method called ridge trace to select the value of the ridge parameter k. This plot shows the ridge regression coefficients as a function of k. When viewing the ridge trace, the analyst picks the value of k for which the regression coefficients have stabilized. Often, the regression coefficients will vary widely for small values of k and then stabilize. We have to choose the smallest value of k possible (which introduces the smallest bias) after which the regression coefficients have seem to remain constant. Hoerl, Kennard and Baldwin (1975), proposed another method to select a single k value given as:

$$\hat{k}_{(HKB)} = \frac{PS^2}{b'_{OLS}\, b_{OLS}} \qquad \ldots\ldots\ldots\ldots\ldots (13)$$

Where P is the number of predictor variables, S^2 is the OLS estimator for σ^2, b_{OLS} is the OLS estimator for the vector of regression coefficients.

Lawless and Wang (1976) have proposed selecting the value of K by using the formula :

$$\hat{k}_{(LW)} = \frac{PS^2}{b'_{OLS}\, X'X\, b_{OLS}} \qquad \ldots\ldots\ldots\ldots\ldots (14)$$

Hoerl and Kennard (1970) suggested the iterative method to estimate the value of K based on the formula:

$$k_{j+1} = \frac{PS^2}{[b_{RR(K_j)}]'[b_{RR(K_j)}]} \qquad \ldots\ldots\ldots\ldots\ldots (15)$$

The first value of k assumed to be zero and hence, $[b_{RR(K_0)}]'[b_{RR(K_0)}] = b'_{OLS}\, b_{OLS}$. Substituting for k_0 in the right hand side of (15) we obtain the first adjusted value k_1 which will be also substituted in the right hand side of equation (15) to obtain the second adjusted value k_2 then continue the iterations until the following inequality have to be satisfied :[5]

$$\frac{k_{j+1} - k_j}{k_j} \leq \varepsilon \ . \qquad \ldots\ldots\ldots\ldots\ldots(16)$$

Where ϵ is small positive number (close to zero). Hoerl and Kennard proposed the value of ϵ to be [5]

$$\varepsilon = 20\left[\frac{\text{trace }(X'X)^{-1}}{P}\right]^{-1.3} \quad \text{.............(17)}$$

In the case of generalized ridge regression Hoerl and Kennard proposed method to select k_i as follows:

$$k_i = \frac{s^2}{\alpha^2_{(OLS)i}} \quad \text{................(18)}$$

Numerical Example

In this section we apply the procedures discussed earlier employing the data obtained from Midland Refineries Company to determine the effect of six factors (explanatory variables X_1, X_2 …X_6) on the productivity of labor (response variable y).The data are given in table (1). Applying the Farrar-Glaubor test given in equation (2), it was shown that the calculated χ^2 is 137.456 while the theoretical value at 15 degree of freedom and 0.05 level of significant is 24.996. Obviously, the calculated value is greater than the theoretical value of the χ^2 which implies that the data suffer from a high degree of multicollinearity. Let us assume that ORR_1 represents the ordinary ridge regression estimator with the ridge parameter obtained by Hoerl-Kennard and Baldwin $\hat{k}_{(HKB)}$, ORR_2 represents the ordinary ridge regression estimator with the ridge parameter obtained by Lawless and Wang $\hat{k}_{(LW)}$ and GRR represents the generalized ridge regression estimator. Applying the formulas in equations (13), (14) and (18) we obtain:

$\hat{k}_{(HKB)} = 0.0075410$, $\hat{k}_{(LW)} = 0.027179$

$\hat{k}_{(GRR)i} = [0.121641, 0.0117446, 0.0051056, 0.053459, 0.002102, 0.083658]$

The computation results of variance inflation factors (VIF) for each explanatory variable, the mean square error (MSE) and the coefficient of determination R^2 for each method are presented in tables (2), (3) and (4) respectively.

Conclusion

In addition to the Farrar-Glaubor test, the large values of VIF in table (2) is another indicator that our dataset suffers from a high degree of multicollinearity [6], since the GRR estimator has smaller MSE and larger R^2 than other two estimators (ORR_1, ORR_2) as it is shown in table (3) and (4), we conclude that the GRR is better than ORR_1 and ORR_2 estimators for remedy the multicollinearity problem in our dataset.

References

1. Obenchain,R.L.(1975),"Ridge analysis following a preliminary test of the shrunken hypothesis",Technometrics, 17. 431-445
2. Gorgees, H.M.(2009),"Using singular value decomposition method for estimating the ridge parameter", Journal of Economics and Administrative science , 53.1-10
3. Hoel, A.E., and Kennard, R.W.(1970), "Ridge Regression Biased Estimation for Nonorthogonal Problem",Technometrics,12 .55-67
4. El-Dereny, M and. Rashwan, N.I.(2011), "Solving Multicollinearity problem Using Ridge Regression Modlels",Int.J.Contemp,Math.Sciences, 6. 585-600 .
5. Drapper, N.R. and Smith, H.(1981), "Applied Regression Analysis", Second Edition, John Wiley and Sons, New York .
6. Pasha,G.R and Ali Shah, M.A.(2004),"Application of ridge regression to multicollinear data", Journal of Research(Science), Bahauddin Zakariya University, Multan, Pakistan , 15. 97-106

Table (1): Effect of six explanatory variables $X_1, X_2 \ldots X_6$ on response variable Y

Y	X_1	X_2	X_3	X_4	X_5	X_6
3193	7	79	305	230	1580	337
3506	8	80	390	266	1590	358
5203	8	81	415	280	1610	416
3118	9	84	425	330	1640	454
10565	9	85	434	368	1642	465
28245	7	137	692	416	1535	470
34701	35	200	759	440	1894	574
33660	3	833	2475	1222	353	533
45240	4	1153	2480	1285	345	733
51157	4	1285	2745	1141	311	873
65085	4	1353	2854	1087	350	878
62893	4	1331	2895	1082	382	908

Table (2): VIF for all variables

Predictor	VIF
X_1	25.279
X_2	366.798
X_3	220.081
X_4	89.442
X_5	884.559
X_6	92.407

Table(3): MSE for each method

MSE	
ORR_1	0.074426
ORR_2	0.093867
GRR	0.069632

Table(4): R^2 for each method

R-Square	
ORR_1	95.94
ORR_2	94.88
GRR	96.2018

توظيف طريقة انحدار الحرف في معالجة مشكلة التعدد الخطي

حازم منصور كوركيس
بشرى عبد الرسول علي
قسم الرياضيات / كلية التربية للعلوم الصرفة (ابن الهيثم) / جامعة بغداد

استلم البحث في:11 ايلول 2012 ، قبل البحث في:15 تشرين الاول 2012

الخلاصة

في هذا البحث قدمنا طرائق عديدة لانحدار الحرف لحل مشكلة التعدد الخطي في أنموذج الانحدار الخطي العام هذه الطرائق تشمل نوعيين من طرائق أنحدار الحرف الاعتيادية (ORR_1, ORR_2) بالاعتماد على طريقة أختيار معلمة الحرف وكذلك تشمل على طريقة أنحدار الحرف العامة (GRR) وقد طبقت هذه الطرائق على مجموعة من البيانات تعاني من مشكلة التعدد الخطي بدرجة عالية وبالاعتماد على معيار متوسط مربعات الخطا (MSE) ومعامل التحديد R^2 لاغراض المفاضلة .تبين لنا ان طريقة أنحدار الحرف العامة (GRR) هي الافضل من الطريقتين الاخرتين .

الكلمات المفتاحية : أنحدار الحرف الاعتيادي،أنحدار الحرف العام، المقدرات المقلصة، تحليل القيمة الشاذة، معامل التحديد