Using the Ridge Regression Procedures to Estimate the Multiple Linear Regression Coefficients

Hazim Mansoor Gorgees, Fatimah Assim Mahdi

1 Baghdad University, College of Pure Science, Ibn Al-Haithem
2 Baghdad University, College of Pure Science, Ibn Al-Haithem

fatima_assim92@yahoo.com, hazim5656@yahoo.com

Abstract. This article concerns with comparing the performance of different types of ordinary ridge regression estimators that have been already proposed to estimate the regression parameters when the near exact linear relationships among the explanatory variables is presented. For this situations we employ the data obtained from tagi gas filling company during the period (2008-2010). The main result we reached is that the method based on the condition number performs better than other methods since it has smaller mean square error (MSE) than the other stated methods.

Keywords: Ordinary ridge regression, Ridge Parameter, Multicollinearity, Condition number, mean square error (MSE).

1. Introduction
Let us consider the classical linear regression model:
\[ y = X\beta + \varepsilon \]  (1)

Where \( y \) is \( (n \times 1) \) vector of response variable, \( X \) is \( (n \times p) \) matrix of explanatory variables, \( \beta \) is \( (p \times 1) \) vector of unknown parameters and \( \varepsilon \) is \( (n \times 1) \) vector of unobservable random errors where \( E(\varepsilon) = 0 \), \( \text{var}(\varepsilon) = \sigma^2 I \)

The problem of multicollinearity occurs when there exist a linear relationship or an approximate linear relationship among two or more explanatory variables. It 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 produced the observed variation in the response variable. Indicators of multicollinearity include a very high correlation among two or more explanatory variables, a very small (near zero) eigen values of the correlation matrix of explanatory variables, a too large condition numbers. and the Farrer- Gloubar test based on the \( \chi^2 \) statistic.

2. Ordinary Ridge Regression Estimator
The most popular method that been proposed to deal with the multicollinearity problem is the ordinary ridge regression estimation method which is a modification of ordinary least squares method to allow biased estimators of regression coefficients. The ridge estimator depend completely upon an exogenous parameter \( (k \text{ say}) \) known as ridge parameter.

For any \( k \geq 0 \), the corresponding ordinary ridge regression estimator denoted by \( b_{ORR} \) is defined as
\[ b_{ORR} = (X'X + k)^{-1}X'y \quad (2) \]

Where the ridge parameter \( k \geq 0 \) is a constant selected by the researcher according to some intuitively plausible criteria put forward by Hoerl and Kennard [1].

3. The choice of Ridge Parameter

The ordinary ridge regression estimator does not provide a unique solution to the multicollinearity problem, instead, it provides a class of solutions which depend upon the ridge parameter \( k \). No explicit optimum value can be found for \( k \). Yet, several stochastic choices have been proposed for this parameter. Some of these choices may be summarized as follows.

Hoerl and Kennard (1970) suggested a graphical method called ridge trace to select the value of the ridge parameter [1]. 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. One has to choose the smallest value of \( k \) (which introduces the smallest bias) after which the regression coefficients seem to be constant.

Hoerl, Kennard and Baldwin (1975) proposed a new approach to select the value of the ridge parameter as follows [1].

\[ k_{HKB} = \frac{\rho s^2}{b_{OLS}' X' X b_{OLS}} \quad (3) \]

Where \( P \) is the number of explanatory variables, \( s^2 \) is the OLS estimator of \( \sigma^2 \) and \( b_{OLS} \) is the ordinary least squares estimator of the vector \( \beta \).

Lawless and Wang (1976) suggested choosing the value of \( k \) by employing the formula [2].

\[ k_{LW} = \frac{\rho s^2}{b_{OLS}' X' X b_{OLS}} \quad (4) \]

Assuming that the regression coefficients vector \( \beta \) has a certain prior distributions, Srivastava (2002) followed Bayesian approach to estimate the ridge parameter [2]. He concluded that is

\[ k_{Bayes} = \text{Max} \{0, \frac{\text{tr}(X'X)}{n-p-1} \frac{b_{OLS}' X' X b_{OLS}}{s^2} - P \} \quad (5) \]

Where \( \text{tr}(XX) \) denote the trace of the matrix \( X'X \) [4].

Gorgees, H. M. and Mahdi, F.A. (2017) utilized the concept of condition number to select the ridge parameter [5].

The condition number is defined to be the ratio of the largest to the smallest singular value of the matrix \( X \) of explanatory variables [5].

The proposed estimator denoted as \( k_{CN} \) is defined to be

\[ k_{CN} = \text{Max} \{0, \frac{\rho s^2}{b_{OLS}' b_{OLS}} - \frac{1}{CN} \} \quad (6) \]

Where \( CN \) refers to the condition number.

Numerical Example

In this section we apply the procedures already discussed making use of the data obtained from Tagi gas filling company through the time period 2008-2016.
We want to assess the effects of four explanatory variable $x_1, x_2, x_3, x_4$ on the response variable $y$, where $y$ represents the annual output of liquid gas cylinders and the explanatory variables $x_1, x_2, x_3, x_4$ refer to craftsmen, administrators, technicians and engineers respectively. The following linear regression model is assumed to specify the relationship between the response variable and the explanatory variables

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \varepsilon_i$$  \hspace{1cm} (7)

Table (1). Values of the explanatory variables $x_1, x_2, x_3, x_4$ and the response variable $y$

|    | $Y$  | $X_1$ | $X_2$ | $X_3$ | $X_4$ |
|----|------|------|------|------|------|
| 1  | 29024876 | 2186 | 490  | 1673 | 312  |
| 2  | 29024876 | 2184 | 464  | 1673 | 325  |
| 3  | 28259383 | 2397 | 510  | 1836 | 357  |
| 4  | 31691496 | 2552 | 544  | 1955 | 380  |
| 5  | 32655027 | 2575 | 549  | 1973 | 383  |
| 6  | 33691061 | 2828 | 604  | 2166 | 421  |
| 7  | 35441678 | 2787 | 593  | 2135 | 415  |
| 8  | 36872615 | 2929 | 624  | 2244 | 436  |
| 9  | 39256145 | 3297 | 702  | 2524 | 490  |

Table (2). Descriptive Statistics

| variable | N | Minimum | Maximum | Mean  | Standard Deviation |
|----------|---|---------|---------|-------|--------------------|
| $Y$      | 9 | 28259383 | 39256145 | 32879684.11 | 3813683.112 |
| $X_1$    | 9 | 2184    | 3297    | 2637.222   | 363.51127   |
| $X_2$    | 9 | 464     | 702     | 564.4444   | 74.2460    |
| $X_3$    | 9 | 1673    | 2524    | 2019.889   | 278.35249  |
| $X_4$    | 9 | 312     | 490     | 391       | 56.16939   |

Table (3). Matrix of Correlation coefficients

|    | $Y$ | $X_1$ | $X_2$ | $X_3$ | $X_4$ |
|----|-----|-------|-------|-------|-------|
| $Y$ | 1.0000 | 0.9578 | 0.9596 | 0.9578 | 0.9498 |
| $X_1$ | 0.9578 | 1.0000 | 0.9951 | 1.0000 | 0.9974 |
Table (4). Analyses of eigenvalues for correlation and matrix condition numbers

| Eigen value | Condition numbers |
|--------------|-------------------|
| 31.882165   | 1                 |
| 0.117604    | 271.0976          |
| 0.000219    | 145580.7          |
| 0.000009    | 3542462.8         |

Table (5). Eigenvectors of correlation matrix

| $X_1$       | $X_2$ | $X_3$       | $X_4$       |
|-------------|-------|-------------|-------------|
| -0.5670     | 0.0527| 1.08514     | 0.5009      |
| 0.4087      | -0.7635| -1.11670    | 0.4988      |
| -0.4183     | 0.0697| 1.10856     | 0.5009      |
| 0.5800      | 0.6398| -1.11163    | 0.4994      |

Table (6). ANOVA in case of OLS

| Source      | d.f | Sum of squares | Mean square | F test |
|-------------|-----|----------------|-------------|--------|
| Regression  | 4   | 7.793412       | 1.948353    | 37.724434 |
| Residual    | 4   | 0.206587       | 0.051646    |        |
Table (7). ANOVA in case of $b_{k=N}$

| Source     | d.f | Sum of squares | Mean square | F test  |
|------------|-----|----------------|-------------|---------|
| Regression | 4   | 7.714816       | 1.928704    | 27.052126 |
| Residual   | 4   | 0.285183       | 0.071295    |         |
| Total      | 8   | 8              |             |         |

Table (8). ANOVA in case of $b_{k=W}$

| Source     | d.f | Sum of squares | Mean square | F test  |
|------------|-----|----------------|-------------|---------|
| Regression | 4   | 7.370573       | 1.842643    | 11.709980 |
| Residual   | 4   | 0.629426       | 0.157356    |         |
| Total      | 8   | 8              |             |         |

Table (9). ANOVA in case of $b_{k=rev}$

| Source     | d.f | Sum of squares | Mean square | F test  |
|------------|-----|----------------|-------------|---------|
| Regression | 4   | 7.304121       | 1.826030    | 10.496257 |
| Residual   | 4   | 0.695878       | 0.173969    |         |
| Total      | 8   | 8              |             |         |

Table (10). ANOVA in the case of $b_{k=CN}$

| Source     | d.f | Sum of squares | Mean square | F test  |
|------------|-----|----------------|-------------|---------|
| Regression | 4 | 7.603587 | 1.900896 | 19.180991 |
| Residual   | 4 | 0.396412 | 0.099103 |
| Total      | 8 | 8         |          |

4. Conclusions and Discussion

Many indicators ensure the presence of multicollinearity problem such as the large values of the correlation coefficients as it is shown in table (3) moreover the very small eigenvalues (near zero) which imply a very large condition number as it is displayed in table (4).

The ANOVA tables displayed that the ordinary ridge regression estimator based on $k_{HKB}^\wedge$ is the best followed by the estimator based on the condition number $k_{CN}^\wedge$ in the sense of residual mean square as it is shown (6),(7),(10),(8),(9).

For the purpose of future works we can use other methods to overcome the multicollinearity problem such as the generalized inverse estimator, Liu estimator, the restricted ridge regression estimator and Jackknife ridge regression estimator.

References

[1] Hoerl, A.E. and Kennard, R.W. (1970); "Ridge Regression Biased Estimation for Non orthogonal problems", Techno metrics, Vol.12, pp 55-67.
[2] Srivastava, M.S. (2002); "Methods of Multivariate Statistics", Wiley, New York.
[3] Pasha, G.R., Shah, M.A. and Ghosia (2004); "Estimation and Analysis of Regression Coefficients When Explanatory Variables are Correlated", Journal of Research (Science), Vol.15, No.1, pp 97-100.
[4] Lindley, D.V. and Smith, A. F. M. (1972); "Bayes Estimators for Linear Models (with Discussions)", J. Roy. Statist. Soc. B 34, pp 1-41.
[5] Gorgees, H. M. and Mahdi, F.A. (2017); "An Alternative Approach for Selecting Ridge Parameter for Ordinary Ridge Regression Estimator", International Journal of Science and Research (IJSR). ISSN 2319-7064.