Comparison of Confidence Interval in the OLS and Ridge Linear Regression Model: A Comparative study via Simulation

Sultana Mubarika Rahman Chowdhury[1], B.M. Golam Kibria[2], Zoran Bursac[1]

ARTICLE HISTORY
Compiled August 30, 2024

[1] Biosatistics Department, Florida International University, Miami, Fl, USA
[2] Department of Mathematics and Statistics, Florida International University, Miami, Fl, USA

ABSTRACT
Write down the abstract here

KEYWORDS
Multiple Linear Regression; Ridge Regression; Multicolinearity

1. Introduction

Multiple linear regression maps the relationship between two or more predictors and dependent variable to a linear equation. The aim is to predict the response variable using the independent variables. If X is a n×p full rank matrix of predictors and Y is a n × 1 vector of response variables the multiple linear regression can be explained as,

$$Y = X \times \beta + \epsilon,$$

where, β is an p × 1 unknown regression paramaters and ϵ is the n × 1 vector of error with mean zero and equal variance.Ordinary Least Square (OLS) method is commonly used to estimate the unknown regression parameter. The OLS estimate in case of linear regression model is defined as follows,

$$\hat{\beta} = (X^T X)^{-1} X^T y.$$

One of the key assumption of the widely used multiple linear regression model is that the predictors need to be independent of each other. Violating the assumption of in-

CONTACT: Sultana Mubarika Rahman Chowdhury. Email: schow034@fiu.edu.

dependence results in an issue known as multicollinearity. Multicollinearity, originally identified by Frisch (1934), is the state in which two or more independent variables show a strong correlation with one another. It causes standard error of the OLS estimator to increase resulting in wide confidence intervals and less reliable results. Identifying and interpreting significant predictors also becomes difficult (Saleh, Arashi, and Kibria 2019). Increasing the sample size, eliminating the highly correlated variables, Principal component analysis are some of the ways to avoid multicollinearity but with the risk of losing either valuable information or interpretability.

As the field of study progressed, researchers developed a number of methods that outperforms OLS in presence of multicollinearity in data by introducing bias to the estimator to gain smaller variance. Ridge regression (Hoerl and Kennard 1970), Lasso (Tibshirani 1996), Stein estimator (Stein et al. 1956), Modified ridge regression, Liu estimator (Liu 1993), Kibria-Lukman estimator (Kibria and Lukman 2020) are some of them. Among these, ridge regression has become one of the most popular methods for dealing with multicollinearity, providing a robust substitute for OLS. However, Ridge regression approach requires precise ridge parameter estimation, because the number of parameters, sample size, degree of correlation, and standard error vary greatly in real-world data. As a results numerous ridge parameters estimators has been suggested researchers till date. Mermi et al. (2024) in their recent paper presented and compared a total of 366 different ridge parameters estimators.

Typically, the performance of various ridge settings are compared to OLS based on their mean square error (MSE). Nonetheless, certain unknown characteristics in the model determines when Ridge Regression estimators outperform Least Squares estimators in terms of MSE (Crivelli et al. 1995). However, that does not tell us how well the corresponding ridge regression model performs in terms of finding out significance of the predictors which is one of the key points in regression analysis.

There are two ways by which statistical significance of the independent variables is determined. One is the method of hypothesis testing and the other one is the method of constructing Confidence Interval (CI). On the basis of power of the hypothesis test, a number of comparative studies have been done for ridge regression with various tuning settings Perez-Melo and Kibria (2020). Nevertheless, when evaluating regression parameters, confidence intervals are preferable to hypothesis testing because they offer a range of plausible values that represent the accuracy, direction and magnitude of the estimate. They provide a more precise grasp of practical significance based on sample size while avoiding the drawbacks of p-values (Nickerson 2000) . In practical fields such as, medical studies investigators are usually interested in determining the size of difference of a measured outcome between groups, rather than a simple indication of whether or not it is statistically significant (Gardner and Altman 1986). Therefore, CI's can be considered as more informative and transparent for making decisions.

This research examines the Confidence Intervals of multiple linear ridge regression settings for a number of shrinkage parameter under identical simulation conditions. The comparison is conducted based on the width of the confidence interval and coverage probability. The research might be useful to evaluate if shrinkage has produced more stable and dependable estimates than ordinary least squares (OLS) regression by evaluating the bias and variance reduction in Ridge Regression. We can also determine which tuning parameter gives a high coverage probability with a comparatively narrow indicating higher precision.

The rest of the paper is organized as follows,

2. Statistical Methodology

2.1. *Ridge Regression Estimators*

2.2. *Confidence Interval*

2.3. *CI for Ridge regression*

3. Simulation Study

4. Application

5. Discussion and Conclusion

Bibliography

Crivelli, Ana, Luis Firinguetti, Rosa Montano, and Margarita Munóz. 1995. "Confidence Intervals in Ridge Regression by Bootstrapping the Dependent Variable: A Simulation Study." *Communications in Statistics-Simulation and Computation* 24 (3): 631–52.

Cule, Erika, Paolo Vineis, and Maria De Iorio. 2011. "Significance Testing in Ridge Regression for Genetic Data." *BMC Bioinformatics* 12: 1–15.

Frisch, Ragnar. 1934. "Statistical Confluence Analysis by Means of Complete Regression Systems." *(No Title)*.

Gardner, Martin J, and Douglas G Altman. 1986. "Confidence Intervals Rather Than p Values: Estimation Rather Than Hypothesis Testing." *Br Med J (Clin Res Ed)* 292 (6522): 746–50.

Gökpınar, Esra, and Meral Ebegil. 2016. "A Study on Tests of Hypothesis Based on Ridge Estimator." *Gazi University Journal of Science* 29 (4): 769–81.

Halawa, AM, and MY El Bassiouni. 2000. "Tests of Regression Coefficients Under Ridge Regression Models." *Journal of Statistical Computation and Simulation* 65 (1-4): 341–56.

Hoerl, Arthur E, and Robert W Kennard. 1970. "Ridge Regression: Biased Estimation for Nonorthogonal Problems." *Technometrics* 12 (1): 55–67.

Kibria, BM Golam, and Adewale F Lukman. 2020. "A New Ridge-Type Estimator for the Linear Regression Model: Simulations and Applications." *Scientifica* 2020 (1): 9758378.

Liu, Kejian. 1993. "A New Class of Biased Estimate in Linear Regression." *Communications in Statistics - Theory and Methods* 22 (2): 393–402. https://doi.org/10.1080/03610929308831027.

Mermi, Selman, Özge Akkuş, Atila Göktaş, and Necla Gündüz. 2024. "A New Robust Ridge Parameter Estimator Having No Outlier and Ensuring Normality for Linear Regression Model." *Journal of Radiation Research and Applied Sciences* 17 (1): 100788.

Nickerson, Raymond S. 2000. "Null Hypothesis Significance Testing: A Review of an Old and Continuing Controversy." *Psychological Methods* 5 (2): 241.

Perez-Melo, Sergio, and BM Golam Kibria. 2020. "On Some Test Statistics for Testing the Regression Coefficients in Presence of Multicollinearity: A Simulation Study."

Stats 3 (1): 40–55.

Saleh, AK Md Ehsanes, Mohammad Arashi, and BM Golam Kibria. 2019. *Theory of Ridge Regression Estimation with Applications*. John Wiley & Sons.

Stein, Charles et al. 1956. "Inadmissibility of the Usual Estimator for the Mean of a Multivariate Normal Distribution." In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, 1:197–206. 1.

Tibshirani, Robert. 1996. "Regression Shrinkage and Selection via the Lasso." *Journal of the Royal Statistical Society Series B: Statistical Methodology* 58 (1): 267–88.