Assessing Robustness of Regularized Regression Models with Applications

Mayooran Thevaraja[1] and Azizur Rahman[2(✉)]

[1] School of Fundamental Sciences, Faculty of Sciences,
Massey University, Palmerston North, New Zealand
[2] Data Science Research Unit (DSRU), School of Computing and Mathematics,
Charles Sturt University, Wagga Wagga, NSW, Australia
azrahman@csu.edu.au

Abstract. In this Big-data and computational innovation era, advanced level analysis and modelling strategies are essential in data science to understanding the individual activities which occur within very complex behavioral, socio-economic and ecological systems. However, the scales at which models can be developed, and the subsequent problems they can inform, are often limited by our inability or challenges to effectively understand data that mimic interactions at the finest spatial, temporal, or organizational resolutions. Linear regression analysis is the one of the widely used methods for investigating such relationship between variables. Multicollinearity is one of the major problem in regression analysis. Multicollinearity can be reduced by using the appropriate regularized regression methods. This study aims to measure the robustness of regularized regression models such as ridge and Lasso type models designed for the high dimensional data having the multicollinearity problems. Empirical results show that Lasso and Ridge models have less residual sum of squares values. Findings also demonstrate an improved accuracy of estimated parameters on the best model.

Keywords: Linear regression · Ridge · Lasso · Cross validation

1 Introduction

In data science, regression methods are utilized for continuous response variable under the supervised learning situation where the aim is to accurately and precisely estimate and predict the outcome, given a set of input variables (Rahman 2017; Das et al. 2018). Typically, such a regression model is trained using a set of objects for which the response is known. The analyst might be interested in the estimated values for the objects in the training set, predicting responses for new objects, identifying which input variables are most important in making good predictions, or inspecting the relationships between these variables (Rahman et al. 2008). Multicollinearity is one of the major issue in the standard regression analysis though (Rahman and Harding 2016).

J. Xu et al. (Eds.): ICMSEM2019 2019, AISC 1001, pp. 401–415, 2020.
https://doi.org/10.1007/978-3-030-21248-3_30