UNIVERSITI TEKNOLOGI MARA | Cawangan Kedah Kampus Sungai Petani

2021 ICMS INTERNATIONAL CONFERENCE ON COMPUTING, MATHEMATICS AND STATISTICS

ROBUST RIDGE REGRESSION APPROACH FOR COMBINED MULTICOLLINEARITY-OUTLIER PROBLEM

Aliah Natasha Affindi, Sanizah Ahmad

Sanizah Ahmad

Faculty of Computer and Mathematical Sciences,

Universiti Teknologi MARA Shah Alam

4 - 5 AUGUST 2021

CONTENTS

Introduction

- **Regression analysis** is often used for parameter estimation using method of **ordinary least squares (OLS)** which offers **good parameter estimates** if all **assumptions are met**.
- However, **if the assumptions are not met** due to **presence** of combined **multicollinearity** and **outliers**, **parameter estimates** may be **severely distorted**.

Introduction

MULTICOLLINEARITY

Multicollinearity is a statistical phenomenon in which two or more variables in a regression model are dependent upon the other variables in such a way that one can be linearly predicted from the other with a high degree of accuracy.

OUTLIERS

Observation that lies an abnormal distance from other values in a random sample from a population.

Objective

To investigate and compare on the **performances** of some robust ridge regression estimators using **simulation study** and **real datasets.**

Robust Ridge Regression Estimators

- Ridge S
- Ridge M
- Ridge MM
- Ridge Least Trimmed Squares (LTS)

Idea of Robust Ridge Regression

Solve Multicollinearity Problem

[Use Ridge Regression]

+

Solve Outliers Problem

[Use Robust Estimator]

=

Solve Multicollinearity and Outlier Problem

[Use Robust Ridge Regression]

- Ridge Regression

- S
- M
- MM
- Least Trimmed Squares (LTS)

- Ridge S
- Ridge M
- Ridge MM
- Ridge Least Trimmed Squares (LTS)

Methodology: Simulation Study

Generate the explanatory variables by using equation

$$x_{ij} = (1-\rho^2)z_{ij} + \rho z_{ij}$$

and levels of multicollinearity

$(\rho = 0.50, \rho = 0.90, \rho = 0.95)$

Generate outliers by introducing two different distributions of error terms:

i. **Laplace distribution** with **mean 0** and **variance 2**

ii. **Cauchy distribution** with **median 0** and **scale parameter 1**

Build model based on different distribution of error term for each **sample size (*n*=25, *n*=50, *n*=100)**

Apply all the **robust ridge e** the model generated

Repeat the simulation for ***m*=1000 times**.

Examine the **performance** of each estimators using the **root mean square error (RMSE)** and **bias**.

The **best method** is the one with **smallest** RMSE and bias values.

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon_i$$

Measurement Criteria

CRITERION	FORMULA
Bias	Bias $= \bar{\beta}_i - \beta_i$ where $\bar{\beta}_i = \frac{\sum_{i=1}^{m} \beta_i}{m}$
Root Mean Square Error (RMSE)	$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{n}\left(\hat{\beta}_i - \beta_i\right)^2}$

where *m* is the number of simulation runs

All analyses were carrie

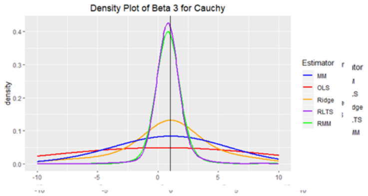
Simulation Results on Laplace Distribution

Method	n	ρ=0.50		ρ=0.90		ρ=0.95	
		Laplace					
		Bias	RMSE	Bias	RMSE	Bias	RMSE
OLS	25	0.0006	0.4105	0.0642	1.6637	0.0185	3.2114
	50	0.0087	0.2754	0.0815	1.1507	0.1099	2.1716
	100	0.0003	0.1875	0.0092	0.7772	0.0229	1.4185
Ridge S	25	0.0898	0.3526	0.0391	0.8296	0.0428	1.5100
	50	0.0537	0.2564	0.0858	0.6269	0.0592	1.0931
	100	0.0223	0.1807	0.0296	0.5020	0.0319	0.7547
Ridge M	25	0.0775	0.3604	0.0258	0.8194	0.0449	1.4647
	50	0.0498	0.2587	0.0833	0.6463	0.0524	1.0387
	100	0.0213	0.1814	0.0264	0.5027	0.0273	0.6906
Ridge MM	25	0.0789	0.3589	0.0278	0.8033	0.0381	1.4202
	50	0.0504	0.2583	0.0839	0.6327	0.0572	1.0086
	100	0.0215	0.1812	0.0280	0.4978	0.0288	0.6785
Ridge LTS	25	0.0965	**0.3493**	0.0241	**0.7738**	0.0472	**1.3622**
	50	0.0553	**0.2559**	0.0916	**0.6023**	0.0661	**0.9818**
	100	0.0229	**0.1805**	0.0248	**0.4842**	0.0429	**0.6642**

Simulation Results on Cauchy Distribution

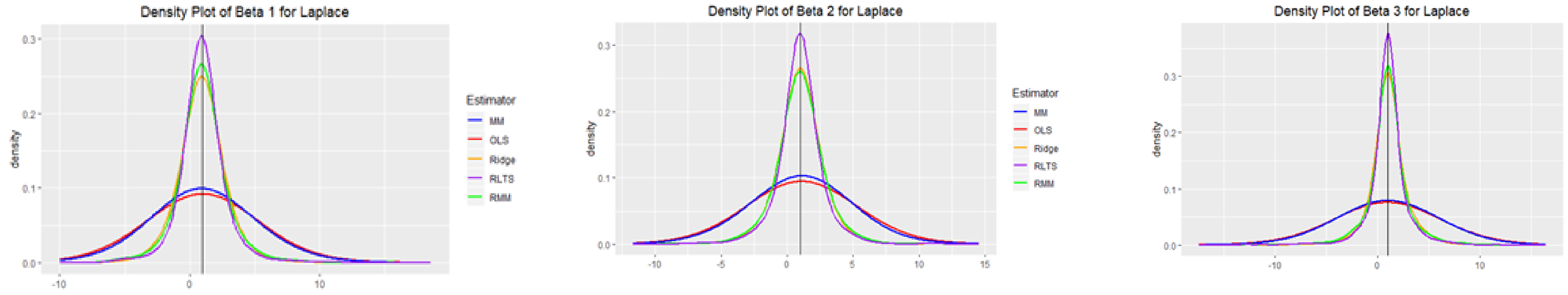
INTERNATIONAL CONFERENCE ON COMPUTING, MATHEMATICS AND STATISTICS

Cauchy							
Method	n	ρ=0.50		ρ=0.90		ρ=0.95	
		Bias	RMSE	Bias	RMSE	Bias	RMSE
OLS	25	3.9811	76.3161	15.7149	301.2479	10.3837	347.9403
	50	0.4014	13.9885	4.6226	321.6129	14.6776	482.7921
	100	0.6616	14.4042	0.4881	75.5959	6.6782	286.1028
Ridge S	25	0.5308	0.7435	0.3195	1.1835	0.1970	1.7309
	50	0.5619	0.7336	0.4015	0.7978	0.2746	1.1347
	100	0.5443	0.7187	0.4484	0.6651	0.3211	0.7507
Ridge M	25	0.4825	0.7733	0.2526	1.3751	0.1996	2.0430
	50	0.5408	0.7441	0.3898	0.8396	0.2706	1.1593
	100	0.5372	0.7205	0.4353	0.6683	0.2771	0.7363
Ridge MM	25	0.5159	0.7371	0.2918	0.9876	0.1920	1.5087
	50	0.5568	0.7343	0.4100	0.7091	0.2711	0.8914
	100	0.5451	0.7176	0.4500	0.6446	0.2995	0.6857
Ridge LTS	25	0.5195	**0.7294**	0.3143	**0.9629**	0.1717	**1.4167**
	50	0.5577	**0.7305**	0.3744	**0.6813**	0.2934	**0.8531**
	100	0.5401	**0.7152**	0.4227	**0.6320**	0.2848	**0.6650**

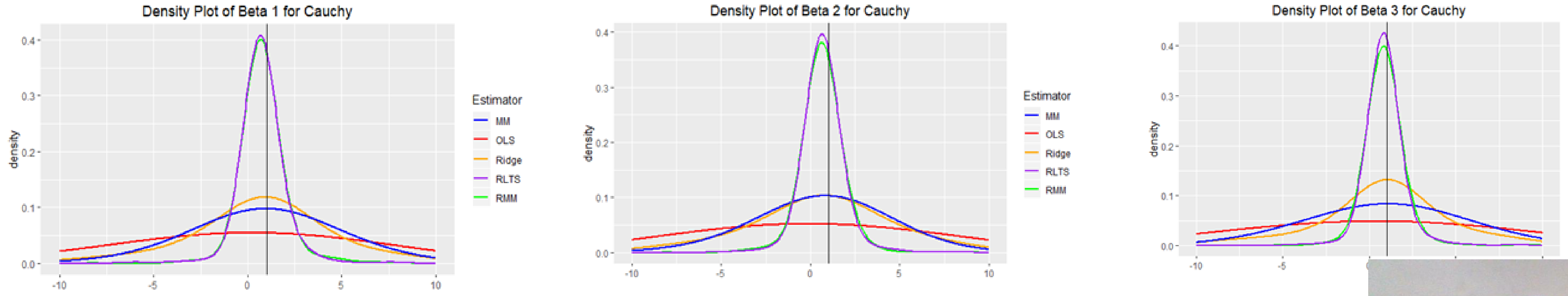


imulation results: Density plots

a)



b)



Density Plots of $\hat{\beta}_1$, $\hat{\beta}_2$ and $\hat{\beta}_3$ for 1000 Simulations for (a) Laplace Distributio

and (b) Cauchy Distribution for ρ=0.95 with n=50.

Application to Real Dataset

Longley dataset (Adegoke, 2016)

This dataset is chosen since the data properties exhibit the interest of study where **both multicollinearity and outliers exist** in the dataset (Cook, 1977; Besley et al., 1980; and Jahufer, 2013).

Longley data consists of **six variables** known as Employment, Prices, Unemployed, Military, GNP and Population Size. GNP is the Gross National Product, employment is the number of people employed, price is the GNP implicit price deflator, unemployed is the number of unemployed, military is the number of people in the armed forces and population size is the non-institutionalized population of persons at age ≥14 years.

Measurement criteria
Standard error (SE) for each estimated parameter

Application on Longley Dataset

Estimate	OLS	Ridge S	Ridge M	Ridge MM	Ridge LTS
$\hat{\beta}_1$	0.0151	-0.0040	-0.0060	-0.0049	-0.0068
SE	0.0849	0.0841	0.0840	0.0840	**0.0840**
$\hat{\beta}_2$	-0.0358	-0.0059	-0.0027	-0.0045	-0.0015
SE	0.0334	0.0276	0.0270	0.0274	**0.0268**
$\hat{\beta}_3$	-0.0202	-0.0157	-0.0153	-0.0155	-0.0151
SE	0.0048	0.0040	0.0039	0.0039	**0.0039**
$\hat{\beta}_4$	-0.0103	-0.0090	-0.0089	-0.0090	-0.0089
SE	0.0021	0.0020	0.0020	0.0020	**0.0020**
$\hat{\beta}_5$	-0.0511	-0.1529	-0.1636	-0.1575	-0.1678
SE	0.2260	0.2167	0.2159	0.2164	**0.2156**
$\hat{\beta}_6$	1.8292	1.3300	1.2776	1.3075	1.2566
SE	0.4554	0.3280	0.3146	0.3222	**0.3092**

Conclusions

- Ordinary least squares (OLS) is not suggested to be used when there exist high multicollinearity and outliers in the data since it may produce high value of mean square error (MSE) and bias which may lead to inaccurate estimation.

- The results of the simulation study was found to be parallel with the result on the real data application where Ridge LTS is the best estimator to be used in the existence of multicollinearity and outliers simultaneously.

THANK YOU

INTERNATIONAL CONFERENCE ON COMPUTING, MATHEMATICS AND STATISTICS

Acknowledgment

The authors wish to thank Universiti Teknologi MARA (UiTM) Shah Alam, Malaysia for the support fund (600-IRMI/REI 5/3 (005/2019)).