BAYESIAN ANALYSIS OF REGRESSION MODEL WITH OUTLIERS AND MISSING DATA: A SIMULATION STUDY

Supported by tetfund TERTIARY EDUCATION TRUST FUND

Oluwadare O. Ojo* and Joseph A. Kupolusi
Department of Statistics, Federal University of Technology Akure, Ondo State, Nigeria
*Corresponding author: daruu208075@yahoo.com

Received: April 9, 2021 **Accepted:** June 16, 2021

Abstract: Outliers and missing value are common problem in applied work. They can lead to inefficient of inferences if they are not properly handled. Bayesian technique had been applied to the two phenomena individually in literature. This work suggested the concept of Bayesian method to handle the problem of outliers and missing data simultaneously in regression model. The suggested Bayesian method was compared with some classical estimators through a simulation study when the regression is characterized by outlier and missing data. The criteria for assessing the performance of these estimators were mean squared error, root mean squared error, mean absolute error, and mean absolute percentage error. Also, in order to evaluate the performance of the model, Akaike and Bayesian information criteria were used. Results from the simulation revealed that Bayesian method of estimation can considerably improve estimation precision.

Keywords: Bayesian technique, missing data, outlier, simulation

JEL classification: C13, C16

Introduction

Outlier is a situation whereby the observed values are usually far from other observations in a data set. There are two kinds of outliers in regression; the first kind of outliers can occur in response variable while second kind of outliers is the one that occur in regressors. Outliers can have a great impact on results of analyses of regression model (Shariff and Ferdaos, 2017). It can also result into heteroscedastcity. The inclusion and exclusion of an observation especially if the sample size is small can substantially alter the results of regression analysis (Gujarati and Porter, 2005). In literature, robust estimators have been specially designed to overcome the problem of outliers since method of Ordinary Least Squares (OLS) is sensitive to small changes in data. Some of the robust estimators are M-estimation developed by Huber (1964), S-estimation by Rousseeuw and Yohai (1984) and MM-estimation introduced by Yohai (1987) among others.

Missing data in regression model simply implies that there is no data value stored for variable in current observation. It can leads into biased estimates and also have a negative impact on statistical power of the model (Mason *et al*., 2010). Missing data can occur in three ways. Randomly missing observations, completely randomly observations, and non-randomly missing observations. Various methods have been proposed to handle the problem of missing data in classical way (Carpenter and Kenward, 2012; El-Sheikh *et al*., 2017). Some recent methods for both outliers and missing data recovery tasks are recorded in the works of Fortuny *et al*. (2015).

Bayesian technique is capable of handling the two problems, since it can take into account of vital information from observed data and uncertainty about the outliers and missing data (Ibrahim *et al*., 2001). Another advantage of Bayesian approach in handling the problems of outlier and missing data is that they are both considered as random variables, whose posterior distributions can be obtained by specifying priors on the parameters. The application of Bayesian technique on missing data is recorded in the works of Swammy and Mechta (1975), Guttman and Menzefrieke (1983), Tanner and Wong (1987), Ibrahim *et al*. (2005), Daniels and Hogan (2008) and recent one Ma and Chen (2018) while notable Bayesian works on outliers is Ekiz (2002), Yuen and Ortiz (2017).

Many methods have been proposed to handle the problem of outlier and missing data individually both in classical and Bayesian ways as mentioned earlier. However, the proposed methods were unable to capture the two problems at the same time. In this work, we applied the proposed Bayesian technique to the problem of outlier and missing data simultaneously and compare this technique with some classical estimators through a simulation study to know the most efficient method.

Materials and Methods

Regression model

Consider a regression model given as:

$$y = x\,\beta + \varepsilon \qquad (1)$$

Where: y is $m \times 1$ vector of responses, x is $m \times k$ matrix of regressors, β is $k \times 1$ vector of regression coefficients, and ε is $m \times 1$ vector of disturbance term of the model assumed to be normally distributed with zero mean and a constant variance σ^2.

The most commonly used technique for solving regression model in (1) is the method of Ordinary Least Squares (OLS). This technique entails minimizing the residuals sum of squares in the model; the estimated $\hat{\beta}$ that minimizes parameter β is given as:

$$\hat{\beta} = (x'\,x)^{-1}\,x'\,y \qquad (2)$$

Bayesian technique for dealing with outliers and missing data

Here, we introduced the method of Bayesian for dealing with both outliers and missing data simultaneously. In Bayesian analysis, the uncertainty about anything unknown can be simply be expressed by the relationship given as:

$$\text{P}\,(\beta|y) \propto \text{P}\,(y|\beta)\,\text{P}\,(\beta) \qquad (3)$$

The quantity, P $(\beta|y)$ is of fundamental interests in this study and entails using the data to learn about parameters in the model given in (1).

Let E denotes the data that is characterized by outliers and missing data.

Then we can now express equation (3) as:

$$\text{P}\,(\beta|E) \propto \text{P}\,(E|\beta)\,\text{P}\,(\beta)$$

The likelihood function of model (1) is given as:

$$\text{P}\,(E|\beta,\Omega^{-1}) \propto |\Omega^{-1}|^{n/2}\exp\,\{-\tfrac{1}{2}\,\Omega^{-1}\,(\beta-\hat{\beta})'M\,(\beta-\hat{\beta}) + vS\,\} \qquad (4)$$

Where: $M = x'x$

$$v = n - r - k + 1$$

$$\hat{\beta} = M^{-1}x'y$$

$$S = \frac{(y-x\hat{\beta})'\,(y-x\hat{\beta})}{v}$$

The quantity $(y - x\hat{\beta})'\ (y - x\hat{\beta})$ can also be written as:

$$(y - x\hat{\beta})'\ (y - x\hat{\beta})$$
$$= (y - xM^{-1}x'y)'\ (y - xM^{-1}x'y)$$
$$= (y - xM^{-1}x'y)'\ (y - xM^{-1}x'y)$$
$$= (y - x(x'x)^{-1}x'y)'\ (y - x(x'x)^{-1}x'y)$$
$$= y'y - y'xM^{-1}x'y - y'xM^{-1}x'y + y'x\ M^{-1}x'xM^{-1}x'y$$
$$= y'y - 2y'xM^{-1}x'y + y'xM^{-1}MM^{-1}x'y$$
$$= y'(I - xM^{-1}x')'(I - xM^{-1}x')y$$

But $(I - xM^{-1}x')'(I - xM^{-1}x')$

$$= [I - xM^{-1}x' - xM^{-1}x' + xM^{-1}x'xM^{-1}x']$$
$$= [I - xM^{-1}x' - xM^{-1}x' + xM^{-1}x']$$
$$= [I - xM^{-1}x'] = \mathrm{R}(x) \qquad (5)$$

N.B: $\mathrm{R}(x)$ is an idempotent matrix

Hence,

$$vS = (y - x\hat{\beta})'\ (y - x\hat{\beta}) = y'\mathrm{R}(x)\,y \qquad (6)$$

Prior distribution reflects the information the researcher has before seeing the data. We assumed a conjugate prior for this study and is given as:

$$\mathrm{P}(\beta^0,\Omega^o) \propto |\Omega^{o^{-1}}|^{\delta/2}\,|\Omega^{o^{-1}}|^{(\mathrm{v}^o - r - 1)/2} \exp[-1/2\ \mathrm{tr}\Omega^{o^{-1}}[\mathrm{v}^o\mathrm{S}^o + (\beta - \beta^0)'\ \mathrm{H}\ (\beta - \beta^0)] \qquad (7)$$

Where: $\zeta = \begin{cases} \zeta(\mathrm{H}) = 0, \text{if H} = 0, \\ 1, \quad \text{otherwise} \end{cases}$

Combining (4) and (7) yields:

$$\mathrm{P}(\beta_*,\ \Omega_*{}^{-1}|\ E) \propto |\Omega^{o^{-1}}|^{(n + \mathrm{v}^o - r - 1 + \zeta)/2} \exp\{-0.5\ \mathrm{tr}(\ \Omega^{o^{-1}})[(v + \mathrm{v}^o)\ g + (\beta^0 - \mathrm{N})'\ (\mathrm{H} + \mathrm{M})\ (\beta^0 - \mathrm{N})]\} \qquad (8)$$

Where: $\mathrm{N} = (M + \mathrm{H})^{-1}\ (M\hat{\beta} + \mathrm{H}\beta^0)$

$$\mathrm{Z} = \frac{(\hat{\beta} + \beta^0)'\ (\hat{\beta} + \beta^0)'}{(H^{-1} + M^{-1})}$$

Then we have;

$$(v + \mathrm{v}^o)\ g = v\ S + \mathrm{v}^o\mathrm{S}^o + \mathrm{Z}\ x$$

If we integrate (8) with respect to Ω^{-1}, we have marginal posterior distribution of β given that data is characterized by both missing data and outliers which is given as:

$$\mathrm{P}(\beta_*|\ E) \propto I_r + \frac{1}{v + \mathrm{v}^o}\{\mathrm{g}^{-1}(\beta^0 - \mathrm{N})'\ (\mathrm{H} + \mathrm{M})\ (\beta^0 - \mathrm{N})\}^{\frac{-(n + \mathrm{v}^o + \zeta)}{2}} \qquad (9)$$

Thus, equation (9) follows a matric-variate t-distribution and will be used for posterior inference to obtain different estimates.

N.B: 0 over and * under represent parameters of prior and posterior distribution, respectively.

Simulation study

In order to assess the performance of the Bayesian procedure with other estimators, we present numerical based on simulated data. Different data sets that is characterized by outliers and missing data were generated based on the regression model given in (1) while necessary criteria will be use to assess the performance of those estimators.

In this study, the regression model that has a relationship between the regressors, disturbance term and response variable will be used and can be simply written as:

$$y = \beta_0 + x_1\,\beta_1 + x_2\,\beta_2 + x_3\,\beta_3 + x_4\,\beta_4 + \varepsilon \qquad (10)$$

The initial values of parameters for the model were set as:

$\beta_0 = 2$, $\beta_1 = 4.5$, $\beta_2 = 10$, $\beta_3 = 0.5$, $\beta_4 = 7$

Prior specifications were set as:

$$\mathrm{S}^o = 0,\ \mathrm{v}^o = 0,\ \beta^0 = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix},\ \Omega^{-1} = I_k$$

The regressors and disturbance term were simulated from a uniform distribution (x_i ~ Unif (0,1)) and normal distribution (ε ~ N(0, 1)), respectively; where i = 1, . . .4. In the simulation, we set the number of observations, sample size to be n = 15, 30, 100, 500 while each of the sample sizes were replicated 5000 times. For each of the datasets, outliers and missing data were introduced. We randomly generate three different percentages of outliers (P) given as:

(a) 0% (no outlier)
(b) 10% outlier
(c) 20% outlier

while missing values were also generated and estimated.

Criteria for evaluation of the estimators are:

- Mean Squared Error (MSE)
- Root Mean Squared Error (RMSE)
- Mean Absolute Error (MAE)
- Mean Absolute Percentage Error (MAPE)

We employed necessary criteria to determine the model performance. The commonly used method proposed by Harvey (1989) and Schwarz (1978) are Akaike Information criterion (AIC) and Bayesian Information Criterion (BIC), respectively.

Results and Discussion

This section presents the results obtained from the simulation by comparing the performances of the estimation methods when the regression model in (13) is characterized by outlier and missing data. Tables 1-4 give the estimates for MSE, RMSE, MAE, and MAPE respectively for different sample sizes. In Tables 5 and 6, AIC and BIC were given to know the model performances for each sample sizes for all the methods. The methods are OLS, M, MM, KNN and Bayesian.

In Tables 1 and 2, Bayesian method of estimation has the smallest MSE and RMSE for all the sample sizes considered in percentages of outliers when a data are missing especially in large sample samples (when n are 100 and 500). KNN method has the highest value of MSE and RMSE for small sample size, that is n=15. All the estimation methods have least values in sample size of 500 for the percentages of outlier compared to other samples for MSE and RMSE. It is obvious that as the percentage of the outlier increases, the values of the estimation method increases.

Table 1: Results of MSE for estimators at different sample sizes

Sample size	Outlier (P)	OLS	M	MM	KNN	Bayesian
15	0	0.4516	0.5643	0.4680	4.4461	0.9943
	10	0.4713	0.7192	0.7910	5.1294	1.9201
	20	1.8419	1.1831	1.9523	5.2197	1.4192
30	0	0.9286	0.9290	0.9286	0.0000	0.0000
	10	1.4215	1.6219	1.6208	0.0191	0.0015
	20	1.5912	2.1951	2.0219	0.1041	0.0159
100	0	0.8573	0.8588	0.8573	0.1310	0.1840
	10	0.9531	0.9514	0.9570	0.0395	0.0400
	20	0.7182	0.6158	0.6171	0.1291	0.1492
500	0	0.9760	0.9761	0.9761	0.1417	0.1058
	10	0.5184	0.5167	0.5164	0.0167	0.0017
	20	0.4156	0.3426	0.3567	0.0271	0.0091

Table 2: Results of RMSE for estimators at different sample sizes

Sample size	Outlier (P)	OLS	M	MM	KNN	Bayesian
15	0	0.6720	0.7512	0.6841	2.1086	0.9972
	10	0.6865	0.8481	0.8894	2.2648	1.3857
	20	1.3572	1.0877	1.3972	2.2847	1.1913
30	0	0.9636	0.9638	0.9638	0.0000	0.0000
	10	0.9259	0.9267	0.9264	0.3620	0.4290
	20	1.2614	1.4816	1.4219	0.3226	0.1261
100	0	0.9259	0.9267	0.9264	0.3620	0.4290
	10	0.9763	0.9754	0.9783	0.1987	0.2000
	20	0.8475	0.7847	0.7856	0.3593	0.3862
500	0	0.9879	0.9880	0.9880	0.3765	0.3252
	10	0.7200	0.7188	0.7186	0.1292	0.0412
	20	0.6720	0.5853	0.5938	0.1646	0.0954

From the results obtained in Tables 3, M method of estimation has the least MAE for sample size of 15 followed by MM method while KNN method has the highest MAE for all the percentages of outlier and when there is missing data. It is apparent that Bayesian method of estimation has least MAE for all other sample sizes considered (n = 30, 100 and 500).

Table 3: Results of MAE for estimators at different sample sizes

Sample size	Outlier (P)	OLS	M	MM	KNN	Bayesian
15	0	0.5387	0.4978	0.5011	1.4012	0.8394
	10	0.6182	0.5161	0.7191	2.8171	0.8231
	20	0.6391	0.5261	0.7812	2.4012	0.8719
30	0	0.7979	0.7965	0.7967	0.0000	0.0000
	10	0.7182	0.7912	0.7915	0.9182	0.8129
	20	0.8129	0.7812	0.7918	0.3226	0.2912
100	0	0.7383	0.7362	0.7366	0.0733	0.0028
	10	0.8123	0.8012	0.8102	0.0812	0.0816
	20	0.8367	0.9471	0.9712	0.0269	0.0012
500	0	0.7924	0.7919	0.7919	0.0673	0.0601
	10	0.7123	0.6912	0.6914	0.0612	0.0539
	20	0.7812	0.7712	0.7129	0.0718	0.0013

Table 4: Results of MAPE for estimators at different sample sizes

Sample size	Outlier (P)	OLS	M	MM	KNN	Bayesian
15	0	0.0474	0.0456	0.0448	0.1116	0.0797
	10	0.0491	0.0451	0.0481	0.1172	0.0791
	20	0.0580	0.0575	0.0564	0.1490	0.0854
30	0	0.0765	0.0764	0.0765	0.0000	0.0000
	10	0.0789	0.0753	0.0759	0.0568	0.0074
	20	0.0791	0.0781	0.0712	0.0182	0.0018
100	0	0.0626	0.0625	0.0625	0.0047	0.0071
	10	0.0691	0.0681	0.0679	0.0051	0.0013
	20	0.7619	0.7213	0.7312	0.0051	0.0025
500	0	0.0699	0.0699	0.0699	0.0047	0.0046
	10	0.0741	0.0731	0.0721	0.0054	0.0051
	20	0.0791	0.0801	0.0791	0.0044	0.0031

In Table 4, Bayesian method of estimation has minimum values for MAPE in most cases of outlier and when there is missing data for all the sample sizes while OLS has the highest MAPE across the sample sizes. For sample size of 15 when there is no outlier and there is missing data, KNN and Bayesian methods have the worst performance.

With the use of AIC and BIC as criteria to know the model performance of the estimation methods as shown in Tables 5 and 6 reveals that Bayesian method has the least AIC and BIC for all the sample sizes considered across percentages of outliers and when there is missing data. However, the KNN performs poorly having the highest AIC and BIC. The values of AIC and BIC increases as the sample sizes increase.

Table 5: Results of AIC for model performance at different sample sizes

Sample size	Outlier (P)	OLS	M	MM	KNN	Bayesian
15	0	1.8919	1.8949	1.8927	2.3652	1.7596
	10	2.7192	2.7219	2.7210	3.1921	2.1197
	20	2.1827	2.1884	2.1843	3.8192	2.0931
30	0	91.5104	91.9334	92.2609	112.7834	91.9017
	10	101.3939	101.5191	102.4849	121.8672	99.3830
	20	111.9490	111.9819	117.3932	133.9284	92.95831
100	0	288.8462	289.1857	289.2414	333.3891	287.1078
	10	219.9401	223.0454	221.4928	312.9415	203.4949
	20	236.8562	240.8739	236.2281	312.9462	201.8384
500	0	1417.640	1417.882	1417.893	1503.488	1416.929
	10	1421.758	1422.0191	1422.4021	1500.817	14012.921
	20	1510.937	1510.958	1510.982	1531.937	1491.935

Table 6: Results of BIC for model performance at different sample sizes

Sample size	Outlier (P)	OLS	M	MM	KNN	Bayesian
15	0	2.0618	2.0648	2.0626	2.5351	1.9295
	10	3.1924	3.2601	3.1919	3.4912	2.1823
	20	2.4019	2.4691	2.4721	2.6591	2.0381
30	0	99.9176	100.3405	100.6681	121.1906	100.3098
	10	111.9594	118.9492	118.9382	133.3939	111.9401
	20	123.4290	129.9420	130.9428	135.9014	104.9481
100	0	304.4772	304.8167	304.8724	349.0201	302.7388
	10	293.8682	294.0398	295.0193	353.2928	283.0296
	20	312.8457	312.4981	313.0127	343.9120	300.1328
500	0	1442.928	1443.170	1443.181	1528.776	1442.216
	10	1431.918	1453.912	1453.293	1473.958	1429.948
	20	1496.938	1501.938	1501.492	1520.392	1472.038

Conclusion

Outliers and missing data is a problem in any empirical work. It can complicate the process of analysis of regression model. Some many methods had been suggested for these two concepts individually. However, this work proposed a Bayesian technique to handle both outliers and missing data simultaneously using normal prior. The proposed Bayesian technique was compared with classical estimators. These classical estimators are Ordinary least squares, M, MM, and K nearest neighbourhood.

Based on the results obtained, Bayesian estimation method outperformed all other estimation methods in the sense of producing least MSE, RMSE, MAE, and MAPE in most cases when there is problem of both outlier and missing data. Bayesian method also has the best model performance with the use of both AIC and BIC as criteria for all cases considered. However, Bayesian method does not really have a good contribution when there is no outlier and but there is missing data especially in small sample sizes. Hence, Bayesian estimation method is the most efficient method and may be recommended for practitioners when tackling the problem of outlier and missing data in regression model.

Acknowledgment
We are grateful to the anonymous reviewers for critical review of this manuscript. Special thanks to Dr. Abonazel for providing part of his computer codes which helps in the production of this manuscript.

References

Abonazel MR 2019. Advance statistical techniques using R: Outliers and missing data. *Annual Conf. on Stat., Comp. Sci. and Operations Res.* Faculty of Graduate Studies for Statistical Research, Cairo University.

Akaike H 1974. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6): 716-723.

Almetwally, E. H and Almongy, H. M. 2018. Comparison between M estimation, S estimation, and MM estimation methods of robust estimation with application and simulation. *Int. J. Mathematical Archive*, 9(11): 55-63.

Carpenter J & Kenward M 2012. Multiple Imputation and its Application. *John Wiley and Sons.*

Daniels MJ & Hogan JW 2008. Missing data in longitudinal studies: Strategies for Bayesian Modeling and Sensitivity Analysis. New York: CRC Press.

Ekiz 2002. A Bayesian method to detect outliers in multivariate linear regression. *Hacettepe,* 31: 77-82.

El-Sheikh A, Abonazel M & Gamil N 2017. A review of software packages for structural equation modeling: A comparative study. *Appl. Maths. and Phy.*, 5(3): 85-94.

Gujarati DN & Porter DC 2009. Basic Econometrics, fifth edition, McGraw Hill, USA.

Huber PJ 1964. Robust version of a location parameter. *Annals of Mathematical Statistics,* 36: 1753–1758.

Fortuny AF, Villaverde AF, Ferter A & Banga JR 2015. Enabling network inference methods to handle missing data and outliers. *BMC Bionformatics,* p. 16.

Fox J 2002. Robust Regression: An R and S-Plus Companion to Applied Regression. *SAGE Publications, Inc.*

Guttman I & Menzefrieke U 1983. Bayesian inference in multivariate regression with missing observation on the response variables. *J. Busin. and Econ. Stat.,* 1: 239 - 248.

Ibrahim JG, Chen MH & Sinha D 2001. Criterion-based methods for Bayesian model assessment. *Statistica Sinica*, 419–443.

Ibrahim JG, Chen MH, Lipsitz SR & Herring AH 2005. Missing-data methods for generalized linear models: A comparative review. *J. Amer. Stat. Assoc.*, 100: 332–346.

Ma Z & Chen G 2018. Bayesian methods for dealing with missing data problems. *J. Korean Stat. Soc.*, 47: 287-313.

Mason AJ, Best N, Plewis I & Richardson S 2010. Insights into the use of Bayesian models for informative missing data. In technical report, London: Imperial College.

Rousseeuw P & Yohai V 1984. *Robust Regression by means of S-estimators*. In Robust and nonlinear time series analysis, 256-272., Springer, New York.

Schwarz GE 1978. Estimating the dimension of a model. *Annals of Statistics*, 6(2): 461 - 464.

Shariff NS & Ferdaos NA 2015. Application of robust ridge regression model in the presence of outliers to real data problem. *J. Phys.: Conf. Series,* 890: 1-6.

Susanti Y, Pratiwi H, Sulistijowati H & Liana T 2014. M estimation, S estimation, and MM estimation in robust regression. *Int. J. Pure and Appl. Maths.*, 91(3): 349-360.

Swamy PAVB & Mehta JS 1975. On Bayesian estimation of seemingly unrelated regression when some observations are missing. *Journal of Econometrics*, 3: 157 – 169.

Tanner MA & Wong WH 1987. The calculation of posterior distributions by data augmentation. *J. Amer. Stat. Assoc.*, 82: 528 - 550.

Yohai VJ 1987. High breakdown-point and high efficiency robust estimates for regression. *The Annals of Statistics,* 15(2): 642-656.

Yuen K & Ortiz GA 2017. Outlier detection and robust regression for correlated data. *Comp. Methods in Appl. Mech. and Engr.*, 313: 632-646.

APPENDICES

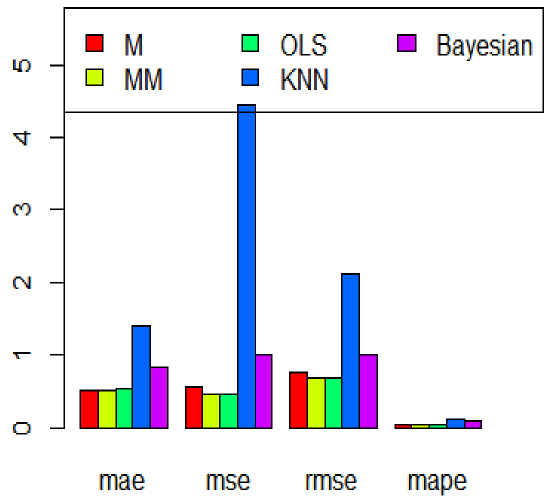

Fig. 1: Bar chart for performance of the methods when sample size, n = 15 for zero outlier and missing value

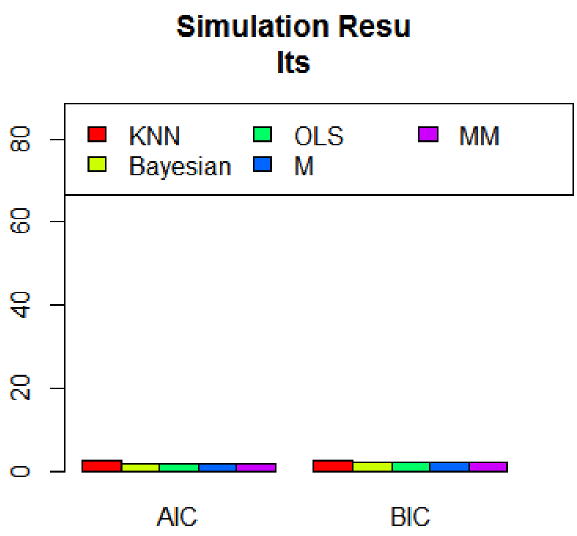

Fig. 2: Bar chart for model performance when sample size, n = 15 for zero outlier and missing value

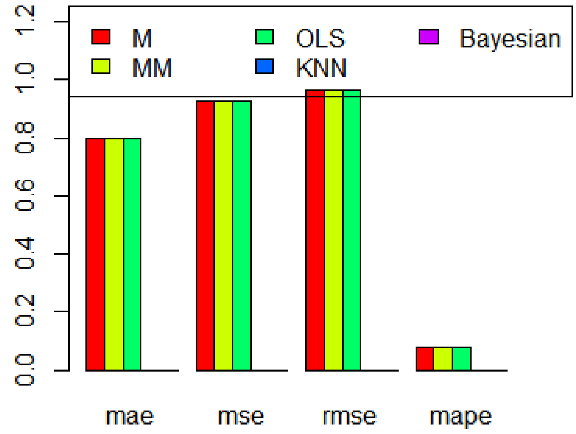

Fig. 3: Bar chart for performance of the methods when sample size, n = 30 for zero outlier and missing value

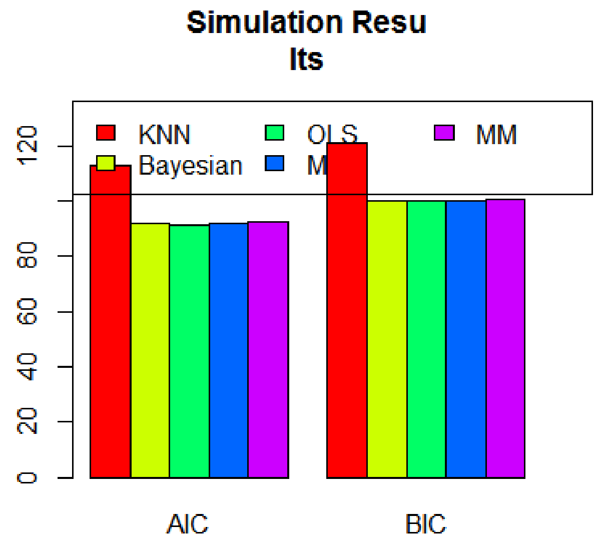

Fig. 4: Bar chart for model performance when sample size, n = 30 for zero outlier and missing value

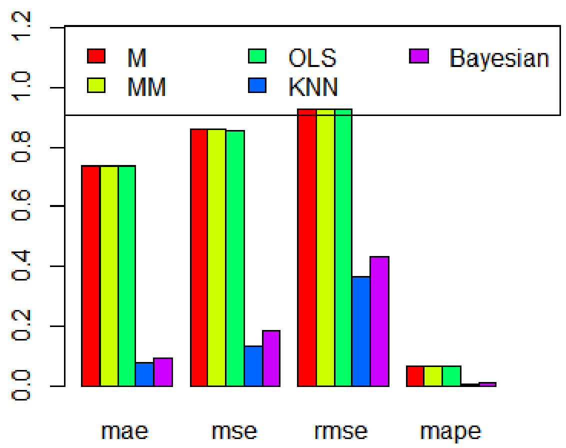

Fig. 5: Bar chart for performance of the methods when sample size, n = 100 for zero outlier and missing value

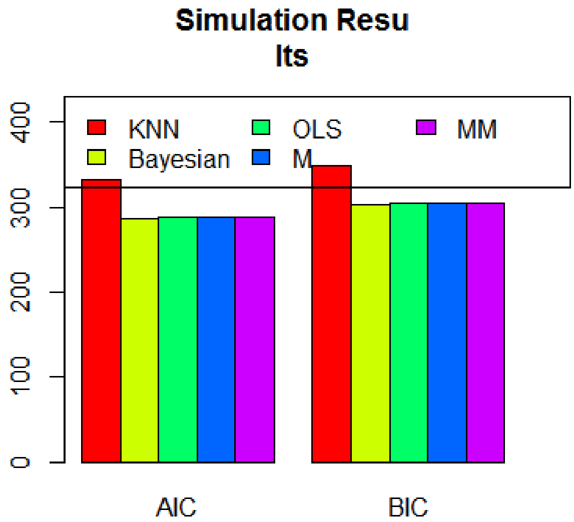

Fig. 6: Bar chart for model performance when sample size, n = 100 for zero outlier and missing value

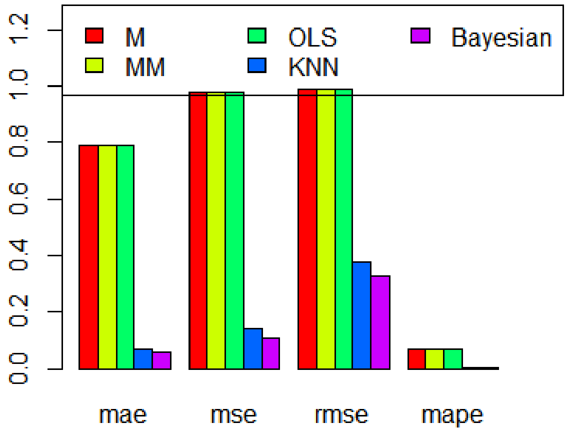

Fig. 7: Bar chart for performance of the methods when sample size, n = 500 for zero outlier and missing value

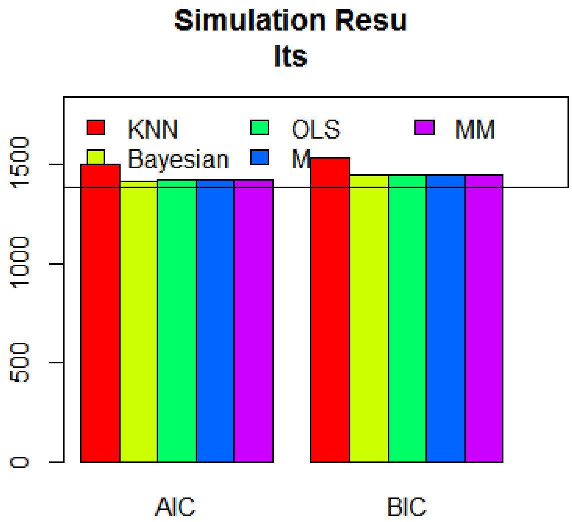

Fig. 8: Bar chart for model performance when sample size, n = 500 for zero outlier and missing value