**Bayesian analysis of regression model with outliers and missing data: A simulation study**

**Ojo O. Oluwadare1 Kupolusi J. Ayodele2**

1,2 Department of Statistics, Federal University of Technology Akure, Nigeria.

Email: 1daruu208075@yahoo.com

2jakupolusi@futa.edu.ng

**Abstract**

Outliers and missing values 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 the 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.

1. **Introduction**

Outlier is a situation whereby an observed value is 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 (see Carpenter and Kenward (2012), El-Sheikh et al. (2017) for details).

Bayesian technique is capable of handling the two problems simultaneously, 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 work on outliers is Ekiz (2002).

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 Bayesian technique to the problem of outlier and missing data simultaneously and compare this technique with some estimators to know the most efficient method.

The rest of the paper is organized as follows. In section 2, we provide the regression model while different estimator methods are discussed in section 3. Bayesian technique for dealing with outliers and missing data was provided in section 4. Simulation study is presented in section 5 while results from the simulation were presented and also discussed in section 6. Section 7 renders the conclusion.

1. **Regression model**

In this section, we will give an overview of linear regression model .

Consider a regression model given as:

 (1)

where is 1 vector of responses, is matrix of regressors, is 1 vector of regression coefficients, and is 1 vector of disturbance term of the model assumed to be normally distributed with zero mean and a constant variance .

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 and the estimated that minimizes parameter is given as:

= ( (2)

1. **Materials and methods**
   1. **M-estimator**

M-estimator is the most common robust regression technique (see Fox (2002), Susanti et al. (2014) for details). It is a generalization of maximum likelihood estimator in terms of location models. This M-estimator for is simply given as:

 (3)

Where 

**MM-estimator**

This estimator was proposed by Yohai (1987) to further overcome the problem of outlier in regression model. It involves the combination of both M and S estimators (see Almetwally and Almongy (2018) for details). The estimator is given as:

=0 (4)

where is the standard deviation of residual of S estimator . The focus of MM estimator is to obtain estimators that have more efficient and a high breakdown value. Hence, the MM estimator of can be simply be defined as:

() = (d) ( - ) (5)

* 1. **K-Nearest Neighbourhood (KNN)**

It is a kind of method used in imputation for missing values. This method entails imputation of missing values given that the number of attributes is the same with the attribute whose values are missing. In KNN, attributes with multiple missing values can be treated and also the correlation structure of the data can be taken into consideration.

**4. Bayesian technique for dealing with outliers and missing data**

This section introduces 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:

P () P () P () (6)

The quantity, P () 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.

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

P () exp { ( ( + } (7)

where, =

=

=

=

The quantity can also be written as:

= +

=

=

But =

=

= ] = R() (8)

N.B: R() is an idempotent matrix

Hence,

= R() (9)

We assumed a conjugate prior which is given as:

P(,) exp[- tr[+(- H (- )] (10)

where

Combining (7) and (10) yields:

P(, ) exp{-0.5 tr[ + + (- N (H + M) (- N)]} (11)

where

N = (+ H (+ H)

Z =

Then we have;

+ = + Z

If we integrate (11) with respect to , we have marginal posterior distribution of given that data is characterized by both missing data and outliers which is given as:

P() {( (H + M) (N) (12)

Equation (12) follows a matric-variate t-distribution.

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

1. **Simulation study**

In order to assess the performance of the Bayesian procedure with other estimators, we present numerical result based on simulated data. Different data sets that are 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:

= + + + + + (13)

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

= 2, = 4.5, = 10, , = 7

Prior specifications were set as:

=0, =0, = , =

The regressors and disturbance term were simulated from a uniform distribution ( ~ Unif (0,1)) and normal distribution ( ~ N(0, 1)) respectively, where = 1, . . .4. In the simulation, we set the number of observations, sample size to be = 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:

1. 0% (no outlier)
2. 10% outlier
3. 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.

1. **Simulation 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 Table 5, 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 are 100 and 500). KNN method has the highest value of MSE and RMSE for small sample size, that is =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 | 0 |
| 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 | 0 |
| 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 ( = 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 | 0 |
| 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 |

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.

**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 | 0 |
| 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 |

With the use of AIC and BIC as criteria to know the model performance of the estimation methods as shown in Table 5 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 and BIC for model performance at different sample sizes**

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | AIC | | | | | BIC | | | | |
| Sample size | Outlier (P) | OLS | M | MM | KNN | Bayesian | OLS | M | MM | KNN | Bayesian |
| 15 | 0 | 1.8919 | 1.8949 | 1.8927 | 2.3652 | 1.7596 | 2.0618 | 2.0648 | 2.0626 | 2.5351 | 1.9295 |
| 10 | 2.7192 | 2.7219 | 2.7210 | 3.1921 | 2.1197 | 3.1924 | 3.2601 | 3.1919 | 3.4912 | 2.1823 |
| 20 | 2.1827 | 2.1884 | 2.1843 | 3.8192 | 2.0931 | 2.4019 | 2.4691 | 2.4721 | 2.6591 | 2.0381 |
| 30 | 0 | 91.5104 | 91.9334 | 92.2609 | 112.7834 | 91.9017 | 99.9176 | 100.3405 | 100.6681 | 121.1906 | 100.3098 |
| 10 | 101.3939 | 101.5191 | 102.4849 | 121.8672 | 99.3830 | 111.9594 | 118.9492 | 118.9382 | 133.3939 | 111.9401 |
| 20 | 111.9490 | 111.9819 | 117.3932 | 133.9284 | 92.95831 | 123.4290 | 129.9420 | 130.9428 | 135.9014 | 104.9481 |
| 100 | 0 | 288.8462 | 289.1857 | 289.2414 | 333.3891 | 287.1078 | 304.4772 | 304.8167 | 304.8724 | 349.0201 | 302.7388 |
| 10 | 219.9401 | 223.0454 | 221.4928 | 312.9415 | 203.4949 | 293.8682 | 294.0398 | 295.0193 | 353.2928 | 283.0296 |
| 20 | 236.8562 | 240.8739 | 236.2281 | 312.9462 | 201.8384 | 312.8457 | 312.4981 | 313.0127 | 343.9120 | 300.1328 |
| 500 | 0 | 1417.640 | 1417.882 | 1417.893 | 1503.488 | 1416.929 | 1442.928 | 1443.170 | 1443.181 | 1528.776 | 1442.216 |
| 10 | 1421.758 | 1422.0191 | 1422.4021 | 1500.817 | 14012.921 | 1431.918 | 1453.912 | 1453.293 | 1473.958 | 1429.948 |
| 20 | 1510.937 | 1510.958 | 1510.982 | 1531.937 | 1491.935 | 1496.938 | 1501.938 | 1501.492 | 1520.392 | 1472.038 |

**7. Conclusion**

Outliers and missing data are problem in any empirical work. They 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, M. R., 2019. Advance statistical techniques using R: outliers and missing data. *Annual* *conference* *on statistics, computer sciences and operations research*, 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*. International Journal of Mathematical Archive*, 9(11), 55-63.

Carpenter, J. and Kenward, M. 2012. Multiple imputation and its application. *John Wiley and sons.*

Daniels, M. J. and Hogan, J. W. 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. and Gamil, N. 2017. A review of software packages for structural equation modeling: A comparative study. *Applied Mathematics and Physics* 5(3), 85-94.

Gujarati, D. N. and Porter, D. C. 2009. Basic Econometrics, fifth edition, McGraw Hill, USA.

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

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

Guttman, I., and Menzefrieke, U. 1983.Bayesian inference in multivariate regression with missing observation on the response variables. *Journal of Business and Economic Statistics,* 1. pp.239 - 248.

Ibrahim, J. G., Chen, M.-H. and Sinha, D. 2001. Criterion-based methods for Bayesian model assessment. *Statistica Sinica*, 419–443.

Ibrahim, J. G., Chen, M.-H., Lipsitz, S. R., and Herring, A. H. 2005. Missing-data methods for generalized linear models: a comparative review. *Journal of the American Statistical Association*, 100, 332–346.

Ma, Z. and Chen, G. 2018. Bayesian methods for dealing with missing data problems, *Journal of the Korean statistical society* 47, 287-313.

Mason, A. J., Best, N., Plewis, I. and 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, G. E., 1978. Estimating the dimension of a model. *Annals of Statistics*, 6 (2): 461- 464.

Shariff, N. S. and Ferdaos, N. A. 2015. Application of robust ridge regression model in the presence of outliers to real data problem. *Journal of Physics: conf series* 890.1-6.

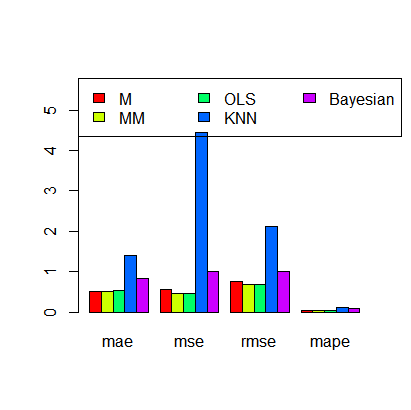
Susanti, Y., Pratiwi, H., Sulistijowati, H., and Liana, T. 2014. M estimation, S estimation, and MM estimation in robust regression. *International Journal of pure and applied Mathematics,* 91:3, 349-360.

Swamy, P. A.V. B., and Mehta, J. S. 1975. On Bayesian Estimation of Seemingly Unrelated Regression when some observations are missing, *Journal of Econometrics*, 3, 157 - 169.

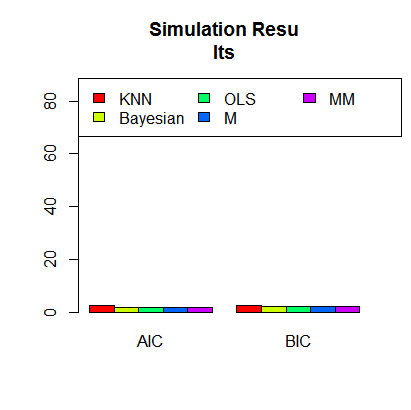
Tanner, M. A., and Wong, W. H. 1987.The calculation of posterior distributions by data augmentation, *Journal of the American Statistical Association*, 82, 528 - 550.

Yohai, V. J. 1987. High breakdown-point and high efficiency robust estimates for regression. *The Annals of Statistics,* 15:2, 642-56.

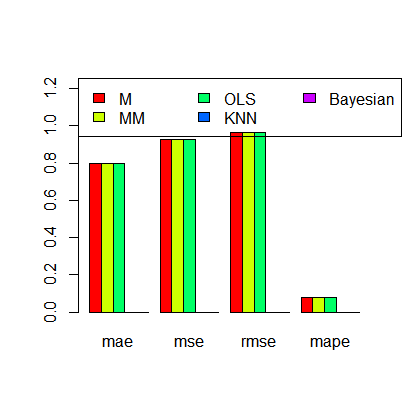
**Appendix**

****

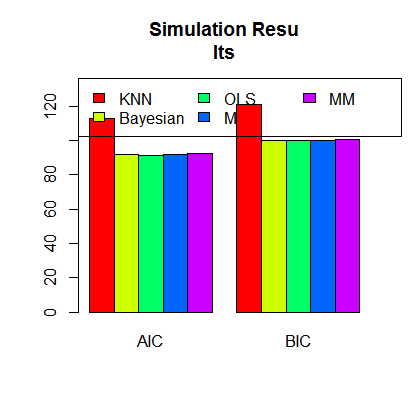
**Figure 1: Bar chart for performance of the methods when sample size, n = 15 for zero outlier and missing value**

****

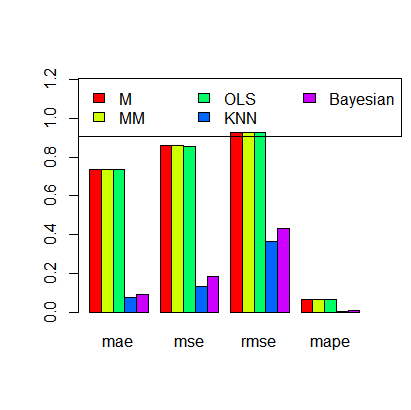
**Figure 2: Bar chart for model performance when sample size, n = 15 for zero outlier and missing value**

****

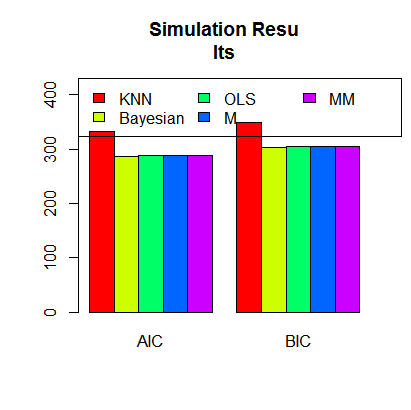
**Figure 3: Bar chart for performance of the methods when sample size, n = 30 for zero outlier and missing value**

****

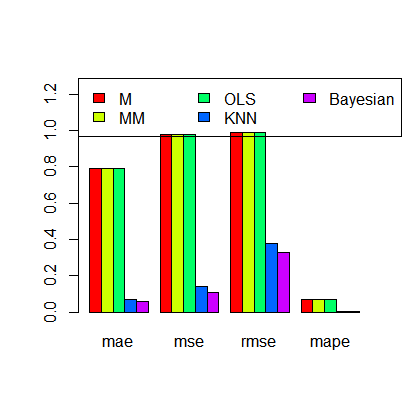
**Figure 4: Bar chart for model performance when sample size, n = 30 for zero outlier and missing value**

****

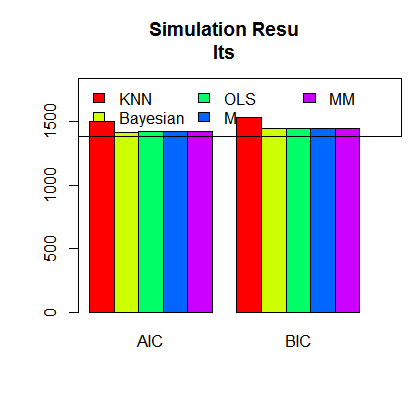
**Figure 5: Bar chart for performance of the methods when sample size, n = 100 for zero outlier and missing value**

****

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

****

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

****

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