Handling Outlier In The Two Ways Table By Using Robust Ammi And Robust Factor

Kurnia Ahadiyah, Alfian Futuhul Hadi, Dian Anggraeni
Graduated School of Mathematics, Mathematics and Natural Science Faculty
Jember University
Kalimantan street on 37, Jember 68121
E-mail: afhadi@gmail.com

Abstract - Robust Regression is a regression methods were used to analyze the data that contains some outliers. This regression is a statistical model were easy to influenced to small changes in the data. This method is often used for data analysis on additive model. In a modeling of two ways table, it has been known AMMI (Additive Main and Multiplicative Interaction) which can be used to analyze the stability of genotypes at several different environments by combining the additive model of main effect and multiplicative model of interaction. AMMI models used for data with normally distribution. AMMI models will against the same problem if there are outliers in two ways table. Because of that problem, it is necessary a robust method in decomposition of interaction matrix including robust SVD and robust PCA. This study analyzed data on two-way tables that contain outliers by using approach of robust SVD and approach of robust PCA. The results of this study on both methods will be compared the goodness of model through the comparison of biplot of each model.

Keywords: *Robust Regression, AMMI, Robust SVD, Robust Factor, Robust PCA.*

INTRODUCTION

In the statistics, regression is the method used to find a causal relationship between the dependent variable (Y) and the independent variable (X). There are a variety of regression that can be used in the analysis of the data depends on the characteristics of the data. Normality of the data is an important thing to consider in linear regression analysis in order to get a good model. However, we often find the data that is nonnormal, such as the data that it contains some outliers.

Robust regression introduced by Andrews (1972) and used for data when the distribution of the error is nonnormal or there are some outliers that affect the model (Ryan, 1997). This method is an important tool to analyze data that is affected by outliers so that the resulting models are robust or resistance against outliers. An estimate that resistance is relatively unaffected by major changes on a small part of data or minor changes on the big part of the data. Robust procedure is intended to accommodate the presence of outliers in the data, as well as abolish the identification of outliers and also automatically in tackling the outliers data (Aunuddin, 1989).

In statistical models, we often find the additive model. Robust Regression was also frequently used or applied on a one-way data that containing outliers when the used model is the additive model. However, we rarely encountered when a data of table in two way at multiplicative models containing outliers analyzed using robust procedures to handling them. In the multiplicative models has been known AMMI (Additive Main and Multiplication Interaction) model. AMMI Model is a model that is often used to analyze the data in the two-way table with main influences of treatment are additive while for the effect of the interaction is modeled with multiplicative models (bilinear). AMMI models can represent a research into the systematic components that contains the major influence (main effect) and the effect of interaction with the multiplicative rates (multiplicative interactions). Random components in this model is assumed to spread normal with constant variance. AMMI models incorporate the additive variance analysis for the primary influence treatment by principal component analysis with bilinear modeling for interaction effects utilizing singular value decomposition (SVD) of the matrix interaction (Mattjik & Sumertadjaya, 2010). However, the interaction matrix susceptible to outliers, while in an effort to obtain superior of genotype properties, the outliers is usually ignored when it could be something that affects the results of the analysis (Mattjik et al., 2011). So it needs other methods used to analyze the data containing outliers when using AMMI models. One method that can be used is a robust method.

Elok (2015) have used the Robust SVD (RobSVD) method on AMMI models. Robust method commonly known as Alternating L1 Regression for singular value decomposition. Ardian (2015) have examined the data that containing outliers (outliers) using robust factor through robPCA approach to obtain the results of the evaluation of handling outliers in the two-ways table data. robPCA (Robust PCA) is based on PCA (Principal Component Analysis) method that solve the data in the presence of outliers. PCA (Principal Component Analysis) is a multivariate analysis that can be used to reduce the dimensions of a variable. In the PCA, the first component is a component that has the greatest variance, while the second component is orthogonal to the first component that maximizes the variance of the point of data that are projected, and the next component in accordance with the eigen vector of covariance matrix. As a result, the first component PCA often leads to the outliers point and does not lead to another observation variance because the variance is smaller. Therefore, the data reduction based on PCA become unreliable if there are outliers in the data (Hubert, Rousseeuw, and Branden, 2005).

In this study, researchers wanted to compare robust method on singular value decomposition through AMMI model and robust factor model using two-way table data.

LITERATUR REVIEW

Robust Regression

Robust regression is an alternative method of least squares method (OLS) to do when the data contains outliers. This method can be used to detect outliers and provide resistant / stable results.

Suppose $X = (x_{ij})$ is a matrix $n \times p$, $y = (y_1, \ldots, y_n)^T$ is a vector of independent variable, and $\theta = (\theta_1, \ldots, \theta_p)^T$ is the parameter vector p which unknown coefficients or components to be estimated. Matrix X is referred to as the design matrix. Known, ordinary linear model can be written as follows :

$$y = X\theta + e$$

with $e = (e_1, \ldots, e_n)^T$ is the n-error vector and the unknown. It is assumed that (any value X) e_i components of e are independent and identically distributed based on $L(\cdot/\sigma)$ distribution. Where σ is a scalar parameter (usually unknown). Often $L(\cdot/\sigma) = \phi(.)$ standard normal distribution with a concentration $\phi(s) = (1/p^2\pi)exp(-s^2/2)$. $r = (r_1, \ldots, r_n)^T$ stated the difference (residual) of n-vector for values expressed by θ and by x_i^T ith row of the matrix X.

ISBN: 978-602-60569-5-5

Ordinary Least Square (OLS) estimate $\hat{\theta}_{LS}$ from θ obtained from the results of $\min_\theta Q_{LS}(\theta)$, where $Q_{LS}(\theta) = \frac{1}{2}\sum_{i=1}^{n} r_i^2$.
By reducing Q_{Ls} against θ is equal to zero, then obtained normal equation,

$$XX^T\theta = X^T y$$

If rank (X) is equal to p then a solution of θ is,

$$\hat{\theta}_{LS} = (X^T X)^{-1} X^T y$$

Estimation of Least Square is the estimate of Maximum Likelihood when $L(\cdot/\sigma)=\phi(.)$. In the case of the usual estimate of scalar parameter σ is

$$\hat{\sigma}_{LS} = \sqrt{\frac{1}{(n-p)} Q_{LS}(\hat{\theta})}$$

The weakness of OLS is this method can only estimate one outlier significantly.

AMMI Model

AMMI models is a multivariate method that is able to explain the average of genotype effect and genotype x environment interaction using PCA (Principal Component Analysis). This model incorporates the effect of additives on the analysis of variance and multiplicative effect on principal component analysis. Bilinear modeling for the interaction effect of genotype by location (γ_{ge}) in this analysis are as follows :

- First of all, arrange the interaction effect in the form of a matrix in which the genotype (rows) × location (column), so that the matrix have the $a \times b$ ordo.
- Furthermore, do the bilinear decomposition of the matrix effects of genotype by environment interactions.

So that the AMMI model can be written as follows:

$$Y_{IJ} = \mu + \alpha_I + \beta_J + \sum_{k=1}^{K} \sqrt{\lambda_k}\gamma_{ik}\,\delta_{jk} + \rho_{IJ} + \varepsilon_{IJ}$$
$$= \mu + \alpha_I + \beta_J + \sqrt{\lambda_1}\gamma_{i1}\delta_{j1} + \sqrt{\lambda_2}\gamma_{i2}\delta_{j2} + \cdots + \sqrt{\lambda_K}\gamma_{iK}\delta_{jK} + \quad \delta_{ij} + \varepsilon_{ij} \qquad (2.3)$$

Where: μ is the general average, α_I is the rows influence (genotype), β_J is the columns influences (environment), $i = 1,2,\ldots,a;\ j = 1,2,\ldots,b;\ k = 1,2,\ldots,m$, with $\sqrt{\lambda_k}$ is the singular value for the k-th component bilinear (λ_k is characteristic root **Z'Z**) $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_K$, γ_{ik} double genotype effect i −th through k-th bilinear component, δ_{jk} double location effect j-th through bilinear component k-th, with constraints :

(1) $\sum_i {\gamma_{ik}}^2 = \sum_j {\delta_{jk}}^2 = 1$, for $k = 1,2,\ldots,m$;

(2) $\sum_i \gamma_{ik}\gamma_{ik'} = \sum_j \delta_{jk}\delta_{jk'} = 0$, for $k \neq k'$;

ρ_{IJ} deviation from bilinier model.

Robust SVD

As the SVD is a least squares procedure, it is highly susceptible to outliers. In the extreme case, an individual cell, if sufficiently outlying, can draw even the leading principal component toward itself. It is therefore desirable to have some way of computing a robust SVD. In SVD method with outliers, the eigenvector can be generated from the covariance matrix. Suppose X is a data matrix which centered on the median, then the equation would be

$${}_nX_p = {}_nX_p^* - ({}_n1_1\ median_{i\leq j\leq p}\ X_{\cdot}^* j)$$

where ${}_n1_1$ is a vector which have value = 1 in each cell and$X_{\cdot}^* j = \left(x_{ij}^*, x_{2j}^*, \ldots, x_{nj}^*\right)$ ia a column vector j-th from matrix X^* for $j = 1,2,\ldots,p$.

The procedure for obtaining eigen value, eigen vector left and right eigen vector iteratively on SVD equation with a robust is called the L1 alternating regression (Warsito, 2009). The L1 Alternating Regression algorithm is:

1. Start with an initial estimate of the leading left eigenvector a_1
2. For each column , $j = 1,2,\ldots,p$, fit the L1 regression coefficient c_j by $\min \sum_{i=1}^{n} |x_{ij} - c_j a_{i1}|$.
3. Calculate the resulting estimate of the right eigenvector $b_1 = \frac{c}{||c||}$ where $||.||$ refers to Euclidean norm.
4. Using this estimate of the right eigenvector, refine the estimate of the left eigenvector. For each row , $i = 1,2,\ldots,n$, fit the L1 regression coefficient d_i by $\min \sum_{j=1}^{p} |x_{ij} - d_i b_{j1}|$.
5. Calculate the resulting estimate of the left eigenvector $a_1 = \frac{d}{||d||}$.
6. Iterate to convergence.

(Hawkins, et. Al. 2001)

Factor Analysis

As for principal components analysis, factor analysis is a multivariate method used for data reduction purpose. Again, the basic idea is to represent a set of variables by a smaller number of variables. In this case they called factors. These factors can be thought of as underlying constructs that cannot be measured by a single variable.

Factor analysis is a technique for analyzing about the dependence of several variables simultaneously with the aim of simplifying of the relations between some of the variables to be a number of factors which is less than the variables. This means, factor analysis can also illustrates the data structure of a research (Suliyanto, 2005).

The factor analysis model can be written algebraically as follows. If you have p variables $X_1, X_2, \ldots, X_p$ measured on a sample of n subjects, then variable i can be written as a linear combination of m factors $F_1, F_2, \ldots, F_m$ where as explained above $m < p$. Thus,

$$X_i = a_{i1}F_1 + a_{i2}F_2 + \cdots + a - im\ F_m + e_i$$

Where the $a_i s$ are the factor loadings (or scores) for variable i and e_i is the part of variable X_i thet cannot be "explained" by the factors. There are three main steps in a factor analysis :

1. Calculate initial factor loadings
2. Factor Rotation
3. Calculation of factor scores

Robust Factor

Factor analysis that uses robust estimator and satisfy the assumptions of factor is called a robust factor analysis. The method used in the robust factor analytic in two-way table is by using estimates Alternating Robust Regression (RAR). RAR is an iterative estimator to get the value of loading scores and estimate scores by alternately or commute.

RESEARCH METHODOLOGY

The data used in this research is secondary data from rice seed research in several districts in Java by consensus of the National Rice in 2008. In this study used 10 varieties of rice and 8 sites experiment with an average of observations (Tons / ha) from three replications. This study uses R program. The steps taken to process the data are:

1. *Outliers identification on Data*
   Identification of outliers done on rice seed research data using car package in R through boxplot and robust method on the effect of rows and columns effect through boxplot.
2. *Modelling data using the robust AMMI model*
   Analysis of AMMI model combine the analysis of variance for the main effect of treatment and principal component analysis on interaction effect matrix. AMMI analysis of the decomposition matrix

of interactions be PCA-PCA (Principal Component Interactions) through a robust approach SVD.

3. *Modelling data using the robust factor model.*
   At this stage, the data is processed by twoway.rob script that has been adapted to program R 2.7.0. In the R program is required weight.wl1 script and the PcaProj script. Twoway.rob script is a script that is used to determine the factor analytic robust models for handling outliers in a two-way table. The purpose of Weight.wl1 script find the value of row weights and column weight of data while the purpose of PcaProj script find the value of the initial score prior iteration.
4. *Compare the model*
   After each model is obtained, the test is done by comparing the model with goodness of fit test of biplot on each model.
5. *Interpretation*
   At this stage represents the results of the algorithm in the form of biplot to estimate the influence of the interaction and interpret models through biplot.

DISCUSSION

Model robust AMMI and models robust factor is implemented on the data seed, which contains 10 genotypes (G1, G2, G3, G4, G5, G6, G7, G8, G9, G10) of rice and 8 sites (L1, L2, L3, L4, L5, L6, L7, L8) with three replications. The three replications were taken the average value because the AMMI model also can modeling the data without replication. The data are as follows:

Table 1. The data

Loc	Rep	Gen	Yield
L1	1	G1	8,004
L1	1	G2	10,832
L1	1	G3	7,532
L1	1	G4	9,774
L1	1	G5	9,003
L1	1	G6	7,513
L1	1	G7	9,661
L1	1	G8	6,700
L1	1	G9	9,838
L1	1	G10	7,074
L2	1	G1	2,653
L2	1	G2	6,650
L2	1	G3	5,878
L2	1	G4	5,429
L2	1	G5	6,181
L2	1	G6	**3,699**
L2	1	G7	6,252
L2	1	G8	4,891
L2	1	G9	1,989
L2	1	G10	2,547
L3	1	G1	9,576
L3	1	G2	10,084
L3	1	G3	9,062
L3	1	G4	9,650
L3	1	G5	6,567
L3	1	G6	9,438
L3	1	G7	9,118
L3	1	G8	6,928
L3	1	G9	8,106
L3	1	G10	9,619
L4	1	G1	6,735
L4	1	G2	7,773
L4	1	G3	7,257
L4	1	G4	7,440
L4	1	G5	7,674
L4	1	G6	6,453
L4	1	G7	7,600
L4	1	G8	6,376
L4	1	G9	6,931
L4	1	G10	7,249
L5	1	G1	6,379
L5	1	G2	5,765
L5	1	G3	6,141
L5	1	G4	5,314
L5	1	G5	5,859
L5	1	G6	6,356
L5	1	G7	7,190
L5	1	G8	7,088
L5	1	G9	7,572
L5	1	G10	5,929
L6	1	G1	8,075
L6	1	G2	8,599
L6	1	G3	7,481
L6	1	G4	7,878
L6	1	G5	9,045
L6	1	G6	7,955
L6	1	G7	8,591
L6	1	G8	8,066
L6	1	G9	8,604
L6	1	G10	7,108
L7	1	G1	7,161
L7	1	G2	7,294
L7	1	G3	7,094
L7	1	G4	7,563
L7	1	G5	7,525
L7	1	G6	7,427
L7	1	G7	8,112
L7	1	G8	8,081
L7	1	G9	8,306
L7	1	G10	7,352
L8	1	G1	7,696
L8	1	G2	7,201
L8	1	G3	7,676
L8	1	G4	7,738
L8	1	G5	7,390
L8	1	G6	7,715
L8	1	G7	7,947
L8	1	G8	7,474
L8	1	G9	6,867
L8	1	G10	7,717

ISBN: 978-602-60569-5-5

From the data above, it necessary to identify outliers in advance to see the outliers in the data. First, the data in boxplot using car package as follows:

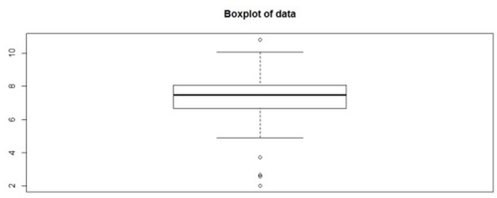

Figure 1

Whereas when using robust factor method obtained outlier identification factor which is reflected in the biplot as follows:

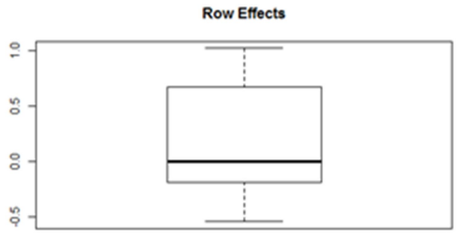

Figure 2

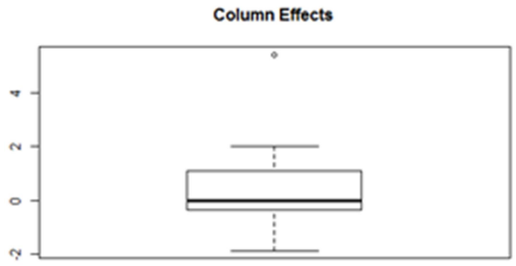

Figure 3

from figures 2 and 3, it appears that the robust method factors may indicate the presence of outliers in the data. Boxplot effect of columns seen that there are outliers in the data.

Modelling with Robust AMMI

In this model, the first analysis performed on additive model using ANOVA. Residual of ANOVA modeled again using singular value decomposition which resistant to outliers is robust SVD.

From robust SVD obtained 7 PCA. From 7 PCA is taken only two components, namely PCA1 and PCA2 to be interpreted in biplot as follows:

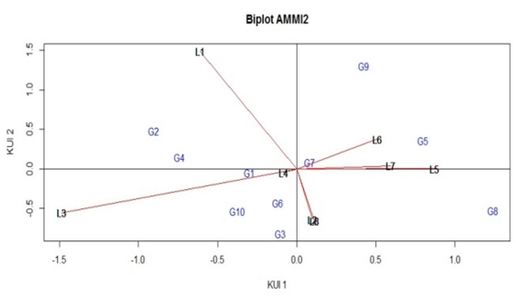

Figure 4

KUI 1 provides information at 48.19% while KUI 2 provides 37.75%. So, the biplot of robust AMMI can interprate the data as big as 85.94% from the real data.

Modelling with Robust Factor

with a robust factor method biplot obtained as follows:

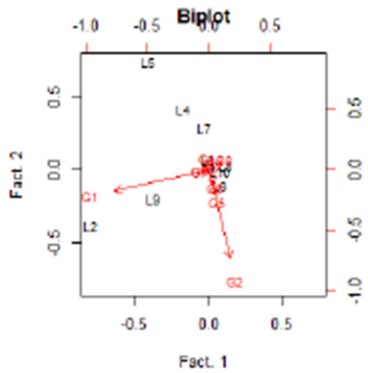

Figure 5

The Biplot has 70,9% of a diversity of the real data.

CONCLUSSION

According the two model, we can obtained a Mean Square Error (MSE) from each model. MSE can make a conclusion in goodness of fit model. Whereas the interpretation of each model, we show in biplot. From biplot, we obtained a percentage value that can make a conclusion about how the diversity of real data. MSE value and percentage biplot value showed in the table below :

	MSE	Biplot
Robust AMMI	0.13191956	85.94%
Robust Factor	0.01057996	59.84%

From that table, we can conclude :

1. The good method for handling the outliers especially in model for prediction is robust factor, because MSE value of robust factor less than MSE value of robust AMMI
2. The good method for handling the outliers especially in interpretation model is robust AMMI, because the percentage of robust AMMI is greater than percentage of robust factor.

REFERENCES

[1]. Aini, E. N. 2015. “Metode Robust Singular Value Decomposition (RSVD) untuk Model AMMI dengan Data Pencilan”. unpublished. Skripsi. Jember : University of Jember.

[2]. Hair, J. F., Anderson, R. E., Tatham, R. L., dan Black, W. C. 1992. *Multivariate Data Analysis with Reading.* 3th Edition. Macmillan Publishing Companyy, Inc.

[3]. Hawkins, D. M., Liu, Li., dan Young, S.S. 2001. Robust Singular Value Decomposition. *National Institute of Statistical Science*.Technical Report Number **122**.

[4]. Huber, P. J. 1986. Projection Pursuit. *The Annals of Statistics*. **13**, 435-475.

[5]. Ryan, T. P. 1997. *Modern Regression Methods*. Canada : John Wiley & Sons,Inc.