ZERO MEAN TRANSFORMATION TECHNIQUE THAT IS NOT EFFECTED BY MISSING OR REMOVED DATA

K.K.L.B. Adikaram[1] and P.A. Jayantha[2]
[1]Computer Unit, Faculty of Agriculture, University of Ruhuna, Sri Lanka
[2] Department of Mathematics, Faculty of Science, University of Ruhuna, Sri Lanka

lasantha@agricc.ruh.ac.lk jayantha@maths.ruh.ac.lk

ABSTRACT: In the process of data transformation, if the mean of the transformed series is zero, such transformation techniques are known as zero mean transformation methods. Mean normalization and standardization are two most common methods that considered as zero mean transformation techniques. Those two methods consider only the dependent variable (*y*) for the transformation but not the independent variable (*x*). Therefore, this approach is suitable for time series that expected to follow $y = c$ relation. Thus, usage of the said methods for time series with missing data that is expected to follow regression other than $y = c$ (*e.g.:* $y = mx + c$), will destroys its original regression and lead to incorrect results. In this paper we represent a zero mean transformation method that transforms any time series into a series that considers both independent and dependent variables. Furthermore, the new technique is independent of the regression of the time series. Furthermore, the proposed technique is resilient to any time series with missing data or removed outliers (without replacement). The results shows that the proposed method is capable of transforming any time series into a series with zero mean despite of the influence of missing or removed outliers.

Keywords: Missing data imputation, Normalization and Standardization, Time series, Transformation techniques, Zero mean

1. INTRODUCTION

Data pre-processing is one critical sub-process of in the data mining or knowledge discovery process [Osborne, 2012]. In the process of pre-processing, *normalization* and *standardization* are two feature scaling techniques. Thus, the major objective of normalization and standardization is almost the same. However, the output of those two techniques is not the same.

In standardization, parameters of different variables will transform to have the same values for selected statistical properties. One formula of such a method, that uses to calculate *standard z score* is given below.

$$y_{new} = \frac{y-\mu}{\sigma}\ , \qquad (1)$$

where y is the original value, y_{new} is the new value, μ is the mean, and σ is the standard deviation. Here, any data set will be transferred to a new data set with 0 mean (μ=0) and unit variance (σ=1)

On the other hand, normalization scales all numeric variables measured on different scales to a notionally common scale such as [0,1]. Scaling is very important when dealing with parameters of different units to have the same scale for a fair comparison between them. One formula of such a method known as *mean normalization* is given below.

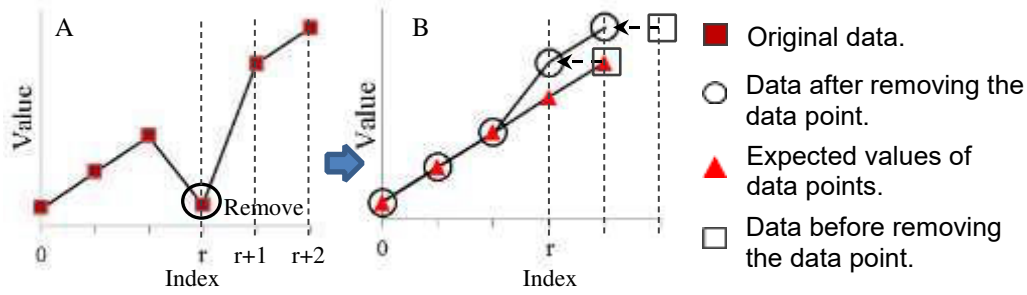
$$y_{new} = \frac{y-\mu}{y_{max}-y_{min}}\ ,$$

ISBN 978-955-627-128-1

(2) where y is the original value, y_{new} is the new value, y_{max} is the maximum of the series y_{min} is the minimum of the series, and μ is the mean. Here, any data set will be transformed into a new data set, where the difference between maximum and minimum terms of the series is one ($max - min = 1$).

Both of these techniques have their drawbacks. If there are outliers in the data set, normalizing data will certainly scale the non-outliers data to a very small interval. When using standardization, new data are not bounded into a range. Also, both the approaches do not consider the independent variable (x) or the sequence of the data. Because of that, if there is no missing or removed data imputation [Pigott, 2001; Nakai, 2011], the original regression will be changed. Figure 1 shows such a situation: the term at index *r* was removed without imputation. After removing the data point at index *r,* data point at index *r+1* is moved to index *r* and so on. Then, the data points after the removed or missing data point will be automatically shifted (Figure 1B) and destroys the original regression [Adikaram et al., 2014].

Figure 1: A.) Data set with a data point that has to be removed (at index r). In plot A, all the data, except the data point at index r, agree with linear regression (y = mx + c). B.) Value auto transformation effect: after removing the data point at index r without a replacement, data point at index r+1 is moved to index r and so on. This will destroy the original regression (y = mx + c).



2. METHODOLOGY

Consider the following relation,

$$y_{new} = y^T - x^T \times \bar{y^T}/\bar{x^T} + c \,, \tag{3}$$

where $y^T = y - y_r$, $x^T = x - x_r$, (x_r, y_r) is any selected reference point, and c is any constant.

From (3)

$\sum y_{new} = \sum(y^T - x^T \times \bar{y^T}/\bar{x^T} + c)$ gives the sum of the terms of transformed series.

$$\begin{aligned}\sum y_{new} &= \sum(y^T - x^T \times \sum y^T/\sum x^T + c)\\ &= \sum y^T - \sum(x^T \times \sum y^T/\sum x^T + c)\\ &= \sum y^T - \sum y^T/\sum x^T \times \sum x^T + \sum c\\ &= \sum c\end{aligned}$$

If the number of terms in the series is *n,* then the mean of the series is given by

$\sum y_{new}/n = \sum c/n.$

ISBN 978-955-627-128-1

When $c = 0$, $\sum y_{new}/n = 0$. (4)

The new method is based on the equation (3). According to the equation (4), when *c =0*, the new transformation method produces a series with zero mean. Thus, the new method can be considered as a zero mean transformation method. The new method was evaluated using empirical data sets and compared the results with *mean normalization* and *standard z score* methods.

3. RESULTS AND DISCUSSION

Consider the data set {10,20,40,50,60,70} which agrees with linear regression (y = 10x) where the third data point is missing. Table 1 and Figure 1 A show the transformation of the data set using *mean normalization* and *standard z score*. Table 2 and Figure 1 B show the transformation of the data set using the new method.

Table 1: Scaling of a data set agrees with linear regression. The third element (30) is missing.

Index (x)	Original Data (y)	Mean normalization	Standard z score
Index or sequence of the data is not engaging in calculations	10	-0.53	-1.37
	20	-0.36	-0.94
	40	-0.03	-0.07
	50	0.14	0.36
	60	0.31	0.79
	70	0.47	1.22
Mean	41.67	0.00	0.00
Standard Dev.	23.17	0.39	1.00
Maximum	70.00		
Minimum	10.00		

Table 2: Scaling of a data set agrees with linear regression using the new method. The third element (30) is missing. However, in the calculations, index of the elements are considered. 6th element is considered as the reference point and c = 0.

Index (x)	Original Data (y)	$x^T = x-x_6$	$y^T = y-y_6$	$y_{new} = y^T - x^T \times$
1	10	-5	-50	0.00
2	20	-4	-40	0.00
4	40	-2	-20	0.00
5	50	-1	-10	0.00
6	60	0	0	0.00
7	70	1	10	0.00
	Sum	-11	-110	0.00
	Mean	-1.83	-18.33	0.00

ISBN 978-955-627-128-1

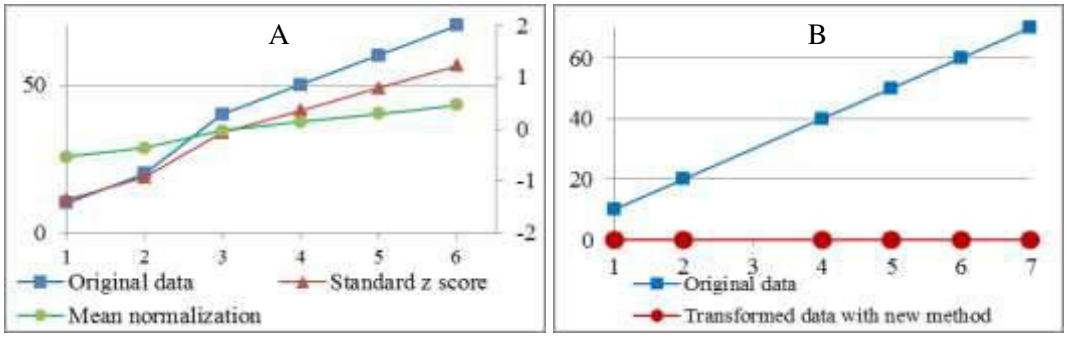


Figure 1: A: Transformation of the data set using mean normalization and standard z score. B: Transformation of the data set using the new method.

The new method was tested with the data sets with both outliers and missing data. Results show that the method produces a zero mean transformed data set for a data set (with outliers and missing data) , which can be expected in real environment (Table 3 and Figure 2).

Table 3: Transformation of a data set agrees with linear regression with missing data and outliers, using the new method. 2nd element is the reference point and c = 0.

Index (x)	Original Data (y)	$x^T = x-x_6$	$y^T = y-y_6$	$y_{new} = y^T - x^T \times y^T / x^T$
1	10	-1	-10	-3.85
2	20	0	0	0.00
4	**60**	2	40	27.69
5	50	3	30	11.54
6	**-10**	4	-30	-54.62
7	70	5	50	19.23
	Sum	13	80	**0.00**
	Mean	2.17	13.33	**0.00**

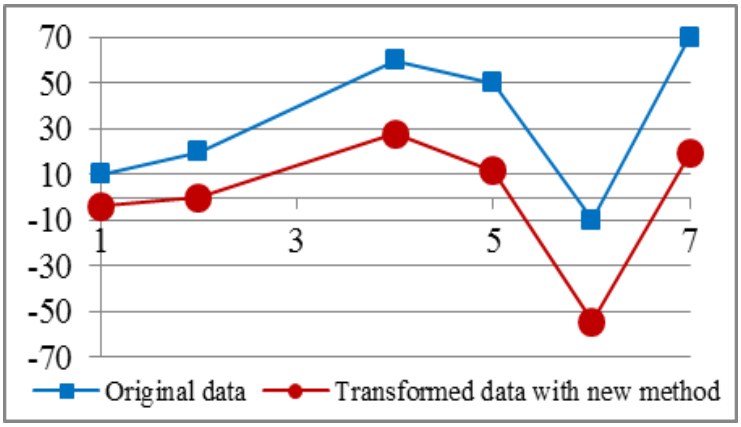


Figure 2: Transformation of a data set agrees with linear regression (Table 3) with missing data and outliers, using the new method.

The reference point in the method is useful in two ways. Reference point can be used to overcome the effect from negative values and it can be used as a baseline point. In the data sets in this paper, reference point is randomly selected. However, selecting a reference point is subjected to the domain requirements. The examples in this paper mainly focusing data sets agree with linear regression. Nevertheless, method can be used with any data set.

ISBN 978-955-627-128-1

4. CONCLUSION

The method can be recommended for normalizing data with initial missing points or outliers and noise. Furthermore, method can be recursively applied after removing outliers and noise without imputing removed data points. Because the method does not need removed or missing data imputation, it will save the computational time.

5. REFERENCES

Adikaram, K.K.L.B., Hussein, M.A., Effenberger, M., Becker, T. (2014). Outlier Detection Method in Linear Regression Based on Sum of Arithmetic Progression. *The Scientific World Journal*, 10.1155/2014/821623.

Michikazu Nakai, W.K. (2011). Review of the Methods for Handling Missing Data in Longitudinal Data Analysis. *Int. Journal of Math. Analysis*, 5(1-4), 1-13.

Osborne, J.W. (2012). *Best Practices in Data Cleaning: A Complete Guide to Everything You Need to Do Before and After Collecting Your Data*. SAGE Publications.

Pigott, T.D. (2001). *A Review of Methods for Missing Data.* Educational Research and Evaluation.

ISBN 978-955-627-128-1