Macedonian pharmaceutical bulletin, **62** (2) 37 - 42 (2016)
ISSN 1409 - 8695
UDC: 616.831-005.1:311.21|(497.7)
Original scientific paper

Effects of data transformation on multivariate analyses in intracerebral hemorrhage

Vladimir Rendevski*[1], Boris Aleksovski[2], Milena Kolevska[2], Dragan Stojanov[3,4], Koco Dimitrovski[5], Ana Mihajlovska Rendevska[6], Vasko Aleksovski[7], Aleksandar Petlickovski[8], Dejan Trajkov[8], Kiro Stojanoski[2]

*[1]University Clinic for Neurosurgery, Medical Faculty, "Ss. Cyril and Methodius" University,
Mother Teresa 17, 1000 Skopje, Republic of Macedonia*
*[2]Faculty of Natural Sciences and Mathematics, "Ss. Cyril and Methodius" University, Arhimedova 3,
1000 Skopje, Republic of Macedonia*
[3]Faculty of Medicine, University of Niš, 18101 Niš, Serbia
[4]Center of Radiology, Clinical Center Niš, 18101 Niš, Serbia
*[5]Institute for Transfusion medicine, Medical Faculty, "Ss. Cyril and Methodius" University,
Mother Teresa 17, 1000 Skopje, Republic of Macedonia*
*[6]University Clinic for Radiology, Medical Faculty, "Ss. Cyril and Methodius" University,
Belgradska bb, 1000 Skopje, Republic of Macedonia*
*[7]University Clinic for Neurology, Medical Faculty, "Ss. Cyril and Methodius" University,
Mother Teresa 17, 1000 Skopje, Republic of Macedonia*
*[8]Institute for Immunobiology and Human Genetics, Medical Faculty, "Ss. Cyril and Methodius" University,
50 Divizija No 6,1000 Skopje, Republic of Macedonia*

Received: October 2016; Accepted: December 2016

Abstract

Multivariate statistical approaches have been increasingly applied in hemorrhagic stroke data analysis. Nevertheless, several aspects regarding their relevance and validity in respect of the application of data transformations have not been studied in details. This paper examines the effects of different data transformations in the standard statistical methods of the multivariate analysis of the intracerebral hemorrhage (ICH) parameters in small group samples. Two different methods for data transformations (log transformation ($\log(X_i)$), square root transformation ($\sqrt{X_i}$))have been carried out. The initial volume of the ICH have been studied using several test for skewness, kurtosis, histogram distribution method and different quartile-quartile (Q-Q) and probability-probability (P-P) plots as criteria for normal distribution. Multivariate analyses for the prediction of the perifocal edema was performed using raw and transformed data. Our results indicate that the data transformation operations should be performed very carefully because different analytical outputs lead to different scientific conclusions.

Keywords: intracerebral hemorrhage, data transformation, perifocal edema, multivariate analysis

* vladimirrendevski@yahoo.com

Abbreviations

ICH - intracerebral hemorrhage; CSS - Canadian Stroke Scale; Q-Qplot - quartile-quartile plot; P-Pplot - probability-probability plot.

Introduction

The intracerebral hemorrhage (ICH) is the deadliest form of stroke worldwide and the leading cause of morbidity and mortality in the developed countries. The age-standardized annual mortality rates (per 100,000) of hemorrhagic stroke has continued to increase (Favate and Younger, 2016; Feigin and Krishnamurthi, 2011).

Multivariate statistical approaches have been widely used for ICH analysis long years now aiming to get new insights in the physiological and biochemical assessments of the stroke. When it comes to ICH, an important characteristic of the statistical data set is that the different variables measured in the same patient are often interrelated or triggered one by another (Brunswick et al., 2012; Castillo et al., 2002; Yan et al., 2016). The usage of statistical methods which assume independence among the variables is frequently not correct and therefore other methodological approaches should be used which take into account the fact that the observed variables are interrelated and they may depict a common pathological mechanism. Simple univariate statistical methods would not be able to describe these interactions and multivariate statistical methods are therefore needed.

A data transformation is usually the first step which is undertaken before performing the multivariate analysis for variety of reasons (ensuring data normality, changing the weights of different variables etc);several different approaches have been proposed for this purpose and some of them are commonly used.In the classical statistical examination, the log transformation is a frequently used method to analyze skewed data and it is one of the most popular transformations used in biomedical research (Feng et al., 2013).

In this paper, several important clinical predictors in patients with ICH have been analyzed and their descriptive, bivariate and multivariate correlations are used in the statistical analysis using binary, categorical variables (age, gender, diabetes mellitus, anatomic localization of the ICH, etc).Different approaches for data transformation of the continuous variables(log transformation, square root transformation) were performed. The initial volume of ICH at admission was analyzed by interpretation of the skewness, kurtosis, the histogram distribution and the generated quartile-quartile (Q-Q) and probability-probability (P-P) plots as methods for examination of normal distribution (Feng et al., 2014; Field, 2009). In addition, multivariate analytical output with transformed and observed data were conducted.The overall aim of the present study was evaluation of the effects of data transformation on the reliability of the results in small data sets from patients with ICH by implementation of comprehensive approach in data analysis.

Methods

A data set of 50 patients with primary, spontaneous, supratentorial ICH recruited from the University clinics of Neurosurgery and Neurology in Skopje (Macedonia) was generated. The information concerning their gender, age, anatomic localization of ICH, blood pressure, blood glucose levels, *Diabetes mellitus*, Canadian Stroke Scale (CSS) scores at admission, initial volume of the ICH and the volume of the peripheral edema five days after ICH were included in the set. The statistical analyses were performed using the statistical software IBM SPSS Statistics22 and Statistica7 (StatSoft©).

The first assumption was that the variables are more or less normally distributed or that the variances are homogenous. Several tests for skewness, kurtosis, histogram distribution method and different Q-Q and P-P plots were generated as criteria for normal distribution (DeCarlo, 1997). Most of the parameters have shown absence of normality of the distribution and two different popular methods for data transformation such as log transformation ($\log(X_i)$) and square root transformation ($\sqrt{X_i}$) were carried out (Bland and Altman, 1996; Box and Cox, 1982).

Results

Data transformations and normal distribution evaluation

The Shapiro Wilk's test have confirmed that the distribution of the values of the initial volume of ICH at admission differed highly significantly from normality ($p = 0.0002$). After log and square root ($\sqrt{X_i}$) transformation, the p values of the performed test have increased, but nevertheless, still confirming absence of normality of the distribution ($p = 0.025$ and $p = 0.042$ respectively). The untransformed (raw) data was characterized as highly positive skewed and slightly kurtotic, resulting in an asymmetric distribution with long tail of the probability density on

Table 1. Skewness and kurtosis coefficients for the initial volume of ICH at admission

Initial volume of ICH at admission (cm³)			
transformation	X_i^*	$\log(X_i)^{**}$	$\sqrt{X_i}^{***}$
skewness	1.030	-0.470	0.303
kurtosis	0.545	-0.770	-0.905

ICH - intracerebral hemorrhage;
*X_i - numerical value of the corresponding concentration in appropriate units;
**$\log(X_i)$ – values gained after logarithmic transformation of the data;
***$\sqrt{X_i}$ – values gained after square root transformation of the data

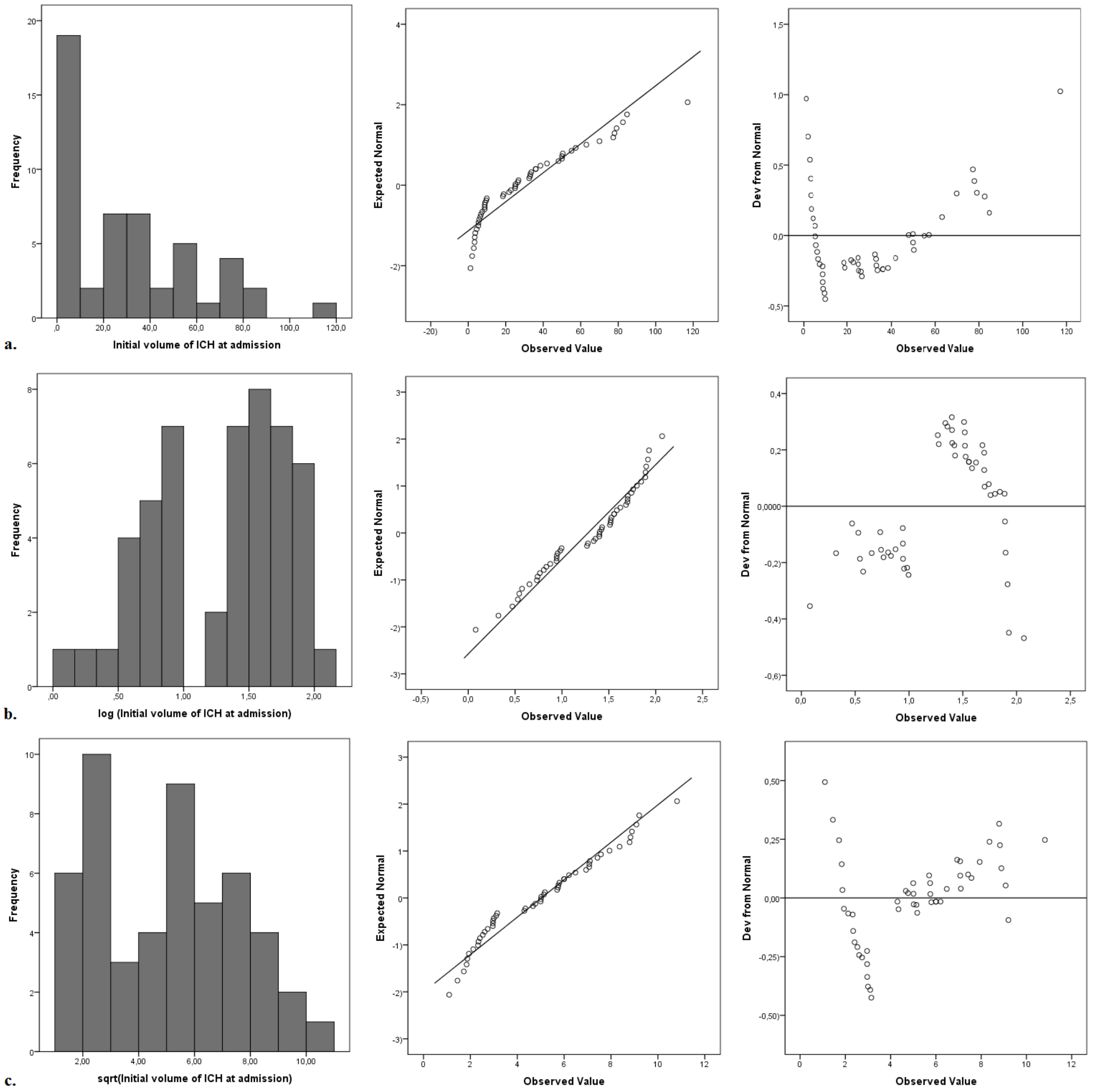

Fig. 1. Normality tests.a. untransformed data (histogram, P-P and Q-Q diagram, respectively). b. log-transformed data (histogram, P-P and Q-Q diagram, respectively). c. square root transformed data (histogram, P-P and Q-Q diagram, respectively).ICH - intracerebral hemorrhage; Q-Q plot - quartile-quartile plot; P-P plot - probability-probability plot; sqrt – square root transformation.

the right side.The logarithmic transformation has reduced the positive skewness, but resulting with slightly negative skewness of the distribution, whereas the $\sqrt{X_i}$transformation has given better results. Nevertheless, both of these transformations have induced an increase of the kurtosis towards higher negative values resulting with slightly platykurtic distribution of the data. The results of the skewness and kurtosis coefficients of the transformed and untransformed data, as well as the tests for normal distribution examination are represented on Table 1 and Figure 1, respectively.

Multivariate statistical analysis

The results using three different models (using raw data, log transformed data and square root transformed data of the variables) for the prediction of the volume of the perifocal edema (five days after ICH) are represented on Table 2.

As it can be noted, the β coefficients and p values for the results of these different multivariate statistical models are approximately similar. All of these three modes have disclosed that the initial volume of ICH at admission is the only significant predictor for the development of the ede-

Table 2. Results from different statistical models with observed and transformed variables (results are obtained by multiple linear regressions with stepwise method)

Parameter	Model 1*		Model 2**		Model 3***	
	β	p value	β	p value	β	p value
Gender	0.129	0.284	0.079	0.410	0.112	0.292
Age	-0.169	0.152	-0.057	0.540	-0.124	0.234
Anatomic localization of ICH	0.062	0.612	-0.001	0.996	0.044	0.673
CSS score at admission	-0.220	0.090	-0.063	0.543	-0.146	0.212
Initial volume of ICH at admission (cm^3)	0.597	0.00001	0.794	7.17×10^{-17}	0.699	1.6×10^{-8}
Glucose (mmol/L)	-0.160	0.203	-0.071	0.454	-0.112	0.296
Diabetes mellitus	-0.157	0.192	-0.048	0.610	-0.102	0.333
Systolic BP (mm Hg)	0.003	0.978	0.005	0.955	0.007	0.945
Diastolic BP (mm Hg)	0.100	0.399	0.089	0.342	0.094	0.368

*Model 1- without data transformation;
**Model 2 - log transformed data;
***Model 3 - square root transformed data.
ICH - intracerebral hemorrhage; CSS - Canadian Stroke Scale; BP – blood pressure.

ma five days after ICH. However, the weight of the β coefficients is different; for instance, when log transformation of the data is used, the influence of the initial volume of the hematoma towards the formation of the edema seems higher when compared to the other models.

Discussion

Ideally, the performance of different type of data transformation in the statistical analyses should be evaluated quantitatively and objectively under the defined circumstances. Unfortunately, an absolute criterion does not exist and it is difficult to estimate whether the proposed method is suitable for our defined scientific purpose.

Despite the usual belief that the data transformation (especially log transformation) can decrease the variability of the data and make data conform more closely to the normal distribution, this is usually not the case. In our study, we have shown that two different data transformation methods did not result in obtaining normal distribution. The transformations have reduced the skewness and the asymmetric probability of the frequencies, but they have resulted in negative kurtosis and have provoked a shift from meso- to platykurtic distribution of the data. The histograms, Q-Q and P-P plots with transformed data have also shown that some transformations can change the type of the distributions and the skewness (see Fig. 1). These results suggest that data transformation is often usefulfor parameters which have positive skewness, where the approximation to a normal distribution can be greatly improved.

The multivariate analyses from the three models have given similar results but, the strength of the β coefficients was different. This can lead to different regression equations and different interpretation of the results. Highest β coefficients were obtained for the log transformed data but, our previous analyses have shown that the $\sqrt{X_i}$ transformation results in a better approximation to normality, which suggests that these β coefficients, although lower, are closer to the scientific reality. It seems that it is important to bear in mind that the transformation of the data can affectthe final estimate(the regression equations form, their characteristics, β coefficients,p values discrepancy from normal distribution etc) as well as the range across the variables and the variance inflation factor (VIF) or, they can reduce the effect of the outliers. The final analytical statistical output can therefore lead to another specific output; in all, before analyzing the data, it seems important to study the structure of the data set because most statistical methods can only give the acceptable answer if the data has the suitable characteristic for the chosen method.

Conclusion

Regardless of the usual belief that the different data transformations(especially log transformation) can decrease the variability of data and make data match more closely to the normal distribution, this is usually not the case. The data transformation operations should be performed very carefully using different statistical models, especially in the case when the most common transformations are performed.In addition, similar multivariate analytical outputs with transformed data were obtained and cross validation studies are needed to define the limits to which extent the transformation model would be acceptable for interpretation of the results.

Acknowledgments

We thank the technical stuff people from the University clinics of Neurosurgery, Neurology and Institute of Transfusion Medicine in Skopje (Macedonia) for their engagement and commitment in the clinical trial.

References

Bland, J.M., Altman, D.G., 1996. The use of transformations when comparing two means. Bmj 312, 1153.

Box, G.E.P., Cox, D.R., 1982. An Analysis of Transformations Revisited, Rebutted. J Am Stat Assoc. 77, 209.

Brunswick, A.S., Hwang, B.Y., Appelboom, G., Hwang, R.Y., Piazza, M.A., Connolly, E.S. Jr., 2012. Serum biomarkers of spontaneous intracerebral hemorrhage induced secondary brain injury. J. Neurol. Sci. 321(1-2), 1-10.

Castillo, J., Dávalos, A., Alvarez-Sabín, J., Pumar, J.M., Leira, R., Silva, Y., Montaner, J., Kase, C.S., 2002.Molecular signatures of brain injury after intracerebral hemorrhage. Neurology 58(4), 624-629.

DeCarlo, L.T., 1997. On the meaning and use of kurtosis. Psychol. Methods 2, 292–307.

Favate, A.S., Younger, D.S., 2016. Epidemiology of Ischemic Stroke. Neurol. Clin. 34, 967–980.

Feigin, V.L., Krishnamurthi, R., 2011. Stroke prevention in the developing world. Stroke 42, 3655–3658.

Feng, C., Wang, H., Lu, N., Chen, T., He, H., Lu, Y., Tu, X.M., 2014.Log-transformation and its implications for data analysis.Shanghai Arch. Psychiatry 26, 105–109.

Feng, C., Wang, H., Lu, N., Tu, X.M., 2013. Log transformation: Application and interpretation in biomedical research. Stat. Med. 32, 230–239.

Field, A., 2009. Discovering statistics using SPSS, third ed. (and sex and drugs and rock 'n' roll). SAGE Publications Ltd, Los Angeles, London, New Delhi, Singapure, Washington DC.

Yan, X.J., Yu, G.F., Jie, Y.Q., Fan, X.F., Huang, Q., Dai, W.M., 2016. Role of galectin-3 in plasma as a predictive biomarker of outcome after acute intracerebral hemorrhage. J. Neurol. Sci. 368, 121–127.

Резиме

Ефекти на трансформација на податоците при мултиваријантна анализа кај интрацеребрална хеморагија

Владимир Рендевски[1], Борис Алексовски[2], Милена Колевска[2], Драган Стојанов[3,4], Кочо Димитровски[5], Ана Михајловска Рендевска[6], Васко Алексовски[7], Александар Петличковски[8], Дејан Трајков[8], Киро Стојаноски[2]

[1]Универзитетска клиника за неврохирургија, Медицински факултет, Универзитет „Св. Кирил и Методиј“,
Мајка Тереза 17, 1000 Скопје, Република Македонија
[2]Природно-математички факултет, Универзитет „Св. Кирил и Методиј“,
Архимедова 3, 1000 Скопје, Република Македонија
[3]Медицински факултет, Универзитет во Ниш, 18101 Ниш, Србија
[4]Центар за радиологија, Клинички центар Ниш, 18101 Ниш, Србија
[5]Институт за трансфузиона медицина, Медицински факултет, Универзитет „Св. Кирил и Методиј“,
Мајка Тереза 17, 1000 Скопје, Република Македонија
[6]Универзитетска клиника за радиологија, Медицински факултет, Универзитет „Св. Кирил и Методиј“,
Белградска бб, 1000 Скопје, Република Македонија
[7]Универзитетска клиника за неврологија, Медицински факултет, Универзитет „Св. Кирил и Методиј“,
Мајка Тереза 17, 1000 Скопје, Република Македонија
[8]Институт за имунобиологија и хумана генетика, Медицински факултет, Универзитет „Св. Кирил и Методиј“,
50 Дивизија 6, 1000 Скопје, Република Македонија

Клучни зборови: интрацеребрална хеморагија, трансформација на податоците, перифокален едем, мултиваријантна анализа

Пристапот на мултиваријантната статистика сè повеќе се применува во анализата на хеморагичниот удар. Како и да е, неколку аспекти во однос на неговата веродостојност и валидност при примената на трансформација на податоците до сега не биле истражени. Овој труд ги истражува ефектите на различните форми на трансформација на податоците при стандардните статистички методи на мултиваријантна анализа кај интрацеребралната хеморагија (ИЦХ) при мала група на испитаници. Два различни метода за трансформација на податоците (log трансформација ($\log(X_i)$) и трансформација со коренување ($\sqrt{X_i}$)) беа извршени. Иницијалниот волумен на ИЦХ беше анализиран користејќи неколку тестови за скјунес, куртозис, хистограмска дистрибуција и различни квартил-квартил (Q-Q) и графици за веројатност (P-P) како критериуми за нормална дистрибуција. Мултиваријантна анализа за предикција на перифокалниот едем беше извршена користејќи нетрансформирани и трансформирани податоци. Нашите резултати укажуваат дека операциите на трансформација на податоците треба да се применуваат мошне внимателно бидејќи различниот аналитички продукт води до различни научни заклучоци.