ENTHUSIASTIC
INTERNATIONAL JOURNAL OF STATISTICS AND DATA SCIENCE
Volume 1, Issue 1, April 2021, pp. 1-6

Mardia's Skewness and Kurtosis for Assessing Normality Assumption in Multivariate Regression

Dewi Wulandari [a,1,*], Sutrisno [a,2], Muhammad Bayu Nirwana [b,3]

[a] Universitas PGRI Semarang, Jl. Dr. Cipto No 24, Semarang 50232, Indonesia
[b] Universitas Sebelas Maret, Jl. Ir. Sutami No. 36, Surakarta 57126, Indonesia
[1] dewiwulandari@upgris.ac.id*; [2] Sutrisno@upgris.ac.id; [3] mbnirwana@staff.uns.ac.id
* corresponding author

ARTICLE INFO

Article history
Received February 20, 2021
Revised March 8, 2021
Accepted April 10, 2021

Keywords
Multivariate skewness
Multivariate kurtosis
Mardia
Multivarate normality assumption
Multivariate regression

ABSTRACT

In Multivariate regression, we need to assess normality assumption simultaneously, not univariately. Univariate normal distribution does not guarantee the occurrence of multivariate normal distribution [1]. So we need to extend the assessment of univariate normal distribution into multivariate methods. One extended method is skewness and kurtosis as proposed by Mardia [2]. In this paper, we introduce the method, present the procedure of this method, and show how to examine normality assumption in multivariate regression study case using this method and expose the use of statistics software to help us in numerical calculation.

1. Introduction

In linear model, normality assumption should be met [3], and we agree that linear regression is of no exception. What if this assumption is not fulfilled? Some previous researchers discussed the unfulfilled normality assumption. On this basis, we take 3 statements about violation of normality assumptions by some researchers: 1) Non-normality distributed variables can distort relationships and significance test [4]; 2) The violation of assumption can have detrimental effects to the result and future directions of any analysis [5]; 3) Violation of the normality assumption may lead to the use of sub optimal estimation, invalid inferential statements, and inaccurate predictions [6].

To satisfy the normality assumption for the sake of dependent variables or response variables as stated in [7], we need to ensure that every value of independent variable, including the corresponding dependent value follows the normal distribution. It shall apply not only to response variables but also to residuals. However, in regression model, it is estimated that using OLS requires assumption of normally distributed error, not the assumption of normally distributed response or predictor variables [8]. Even though we object to that statement, because we are yet to study that in-depth, in this paper we demonstrate only the normality assumption for residuals or errors.

There are many methods to assess normality assumption. One of those methods is skewness and kurtosis test. Kim [9] stated that there is no current gold standard method to assess normality of data. Saphiro-Wilk test and Kolmogorov-Smirnov test are regarded as an unreliable tests for large samples, while Skewness and kurtosis test may be relatively correct in both small and large sample. The importance of normality assumption not only lies in univariate models, but also in multivariate

p-ISSN 2798-253X
e-ISSN 2798-3153

models. Assumptions for a multivariate regression analysis are similar to the assumptions under the univariate regression, but they are extended to a multivariate domain [10]. Multivariate skewness and kutosis proposed by Mardia is one of the extended methods. Romeu and Ozturk [11] did 10 tests for multivariate normality using skewness and kurtosis and the result showed that Mardia's skewness and kurtosis were the most stable and reliable. In this paper, we demonstrate the assessment procedure of the normality assumption in multivariate regression using data and multivariate regression model by Khasanah [12].

2. Methods

Using literature study method in this paper, we collected references, searched for the urgency of normality assumption in regression analysis, examined the procedure of skewness and kurtosis test to do assessment, applied it to the multivariate regression case, and showed how to use the statistic software to help in numerical calculations.

3. Results and Discussion

In this section, we deliver the multivariate regression, skewness and kurtosis test, and the use of statistics software in numerical calculation.

3.1. Multivariate Regression

Simple linear regression model is stated by (1).

$$Y_i=\beta_0+\beta_1 X_i+\varepsilon_i \quad (1)$$

where Y_i is the i^{th} response, β_0 is the intercept, β_1 is the regression coefficient, X_i is the i^{th} predictor and ε_i is the i^{th} random error.

If we have more than one predictors, we will have a multiple linear regression model, which is expressed by (2).

$$Y_i=\beta_0+\beta_1 X_{1i}+\ldots+\beta_k X_{ki}+\varepsilon_i \quad (2)$$

If we have a multiple linear regression with more than one response variables, we will have a multivariate multiple regression, which is stated by (3).

$$\mathbf{Y}=\mathbf{X\beta}+\boldsymbol{\varepsilon} \quad (3)$$

where $\boldsymbol{Y}$ is n×k response matrix, $\mathbf{X}$ is n×(p+1) predictor matrix, $\boldsymbol{\beta}$ is (p+1)×k regression parameter matrix and $\boldsymbol{\varepsilon}$ is n×k random error matrix. The model obtained after an estimation is presented in (4), where $\widehat{\mathbf{Y}}$ is n×k estimated response matrix and $\widehat{\boldsymbol{\beta}}$ is (p+1)×k estimated regression parameter matrix. The least square estimator of $\boldsymbol{\beta}$ is $\mathbf{(X^tX)^{-1}X^tY}$.

$$\widehat{\mathbf{Y}}=\mathbf{X}\widehat{\boldsymbol{\beta}} \quad (4)$$

In this paper, we used the data and multivariate regression model from the case stated in the thesis of Khasanah [12] to do the simulation. In this case, the response variables are the mathematics cognitive and affection score in grade 8 (Y_1 and Y_2 respectively). Khasanah [12] used 4 predictors: the mathematics cognitive and affection score in grade 7 (X_1 and X_2 respectively), parent's salary per month (X_3) and teachers' experience (X_4), i.e. the period of the teachers' teaching carrier. There are total 148 observations used.

The estimated multivariate models are shown in (5) and (6).

$$Y_1=3.00502+0.86625X_1+0.08785X_2+1.22701X_3-0.03247X_4 \quad (5)$$

$$Y_2=1.95182+0.04743X_1+0.91259X_2+0.36732X_3-0.11298X_4 \quad (6)$$

Formula (5) and (6) can be written as (4), where:

$$\widehat{\mathbf{Y}}=\begin{bmatrix} Y_{11} & Y_{12} \\ \vdots & \vdots \\ Y_{n1} & Y_{n2} \end{bmatrix};\ \mathbf{X}=\begin{bmatrix} 1 & X_{11} & \cdots & X_{14} \\ 1 & X_{21} & \cdots & X_{24} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & \cdots & X_{n4} \end{bmatrix};\ \widehat{\boldsymbol{\beta}}=\begin{bmatrix} 3.00502 & 1.95182 \\ 0.86625 & 0.04743 \\ 0.08785 & 0.91259 \\ 1.22701 & -0.36732 \\ 0.03247 & -0.11298 \end{bmatrix}$$

From the estimated model, we determine the residual matrix by subtracting $\widehat{\mathbf{Y}}$ from $\mathbf{Y}$.

p-ISSN 2798-253X
e-ISSN 2798-3153

3.2. Skewness and Kurtosis

In univariate case, skewness is a measure of the asymmetry of the distribution of a variables. The zero value of skewness shows that the data follow normal distribution and any symmetric data should have skewness near zero. Positive value of skewness indicates that the tail on the right side is longer than the left side and vice versa. These situations are presented in Fig 1.

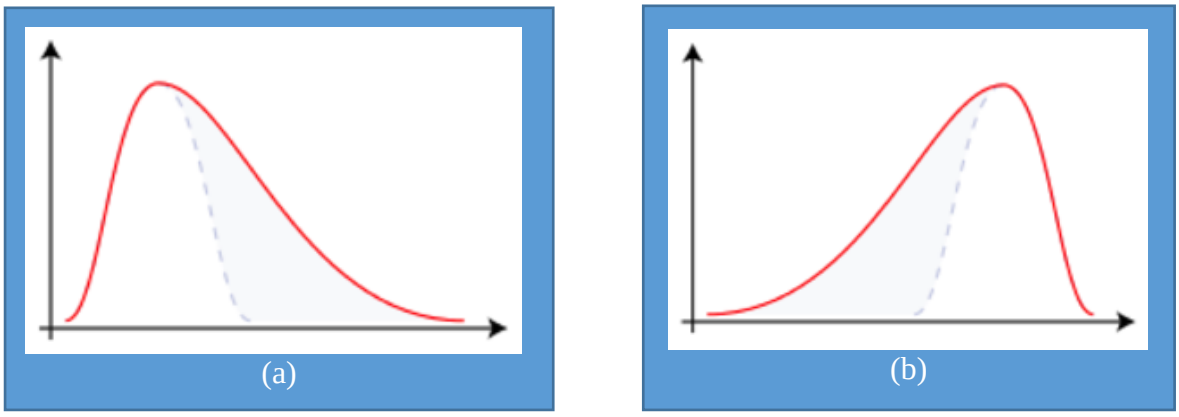

Fig. 1. a) Positive Skewness b) Negative Skewness

Kurtosis is measure of a peakedness of a distribution. The data which have high kurtosis value will have heavy tails and the kurtosis is known as leptokurtic kurtosis. Meanwhile, kurtosis with the opposite condition is defined as platykurtic kurtosis. Fig. 2 shows us the peakedness of a distribution.

Kim [9] applied a z-test for univariate normality test using skewness and kurtosis. The statistic value are expressed in (7) and (8).

$$z_S = \frac{\text{Skew Value}}{SE_{skewness}} \tag{7}$$

$$z_K = \frac{\text{Excess Kurtosis}}{SE_{excesskurtsis}} \tag{8}$$

where z_S is a z-score for skewness and z_K is a z-score for kurtosis. When given observations data $(X_1,X_2,X_3,\ldots,X_n)$ the skew value formula is $(\sum_{i=1}^{n}(X_i-\bar{X})^3/n)/s$, $SE_{skewness}$ formula is $\sqrt{(6n(n-1))/(n-2)(n+1)(n+3)}$, the excess kurtosis formula is $\left(\frac{\sum_{i=1}^{n}(X_i-\bar{X})^4}{n}/s^4\right)$-3, and the $SE_{exceskurtosis}$ formula is $\sqrt{6n/(n-2)(n-3)(n+3)(n+5)}$. According to Kim [9], the critical value for rejecting the null hypothesis needs to be differentiated based on the sample size:

1. For small samples ($n<50$), if the absolute z-score for skewness or kurtosis is less than 1.96, corresponding to an alpha level of 0.05, then do not reject the null hypothesis. It means that the sample is normally distributed.
2. For a medium-sized sample ($50 \leq n \leq 300$), do not reject the null hypothesis at an absolute z value under 3.29, corresponding to an alpha level of 0.05, and conclude the sample distribution is not normally distributed.
3. For sample sizes greater than 300 rely on the histogram and absolute values of skewness and kurtosis regardless of the z value. Either an absolute skew value greater than 2 or an absolute kurtosis greater than 7 can be used as a reference value to determine substantial non-normality.

p-ISSN 2798-253X
e-ISSN 2798-3153

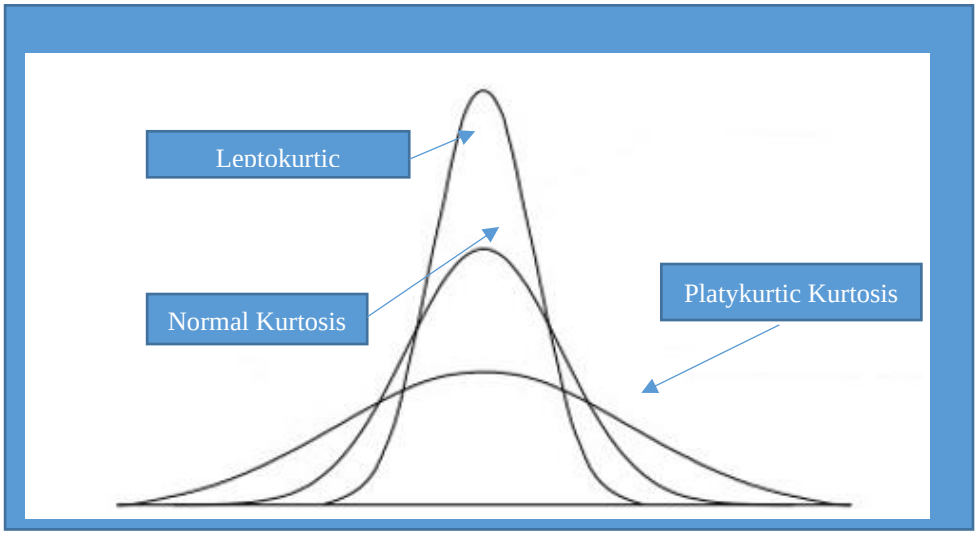

Fig. 2. Kurtosis

The procedure for multivariate skewness and kurtosis test is discussed in [13]. The null hypothesis stated that the sample is from the normal distribution population. The recommended significance level is 5%. In this section, we determine the value of multivariate skewness and kurtosis and then the critical value and decision is discussed in section 3.3.

In the case as determined in section 3.1., we have bivariate residuals. The multivariate skewness is expressed in (9) by supposing that $\mathbf{X}_i^t=(X_{1i},X_{2i},\ldots,X_{pi})$, i=1,2,…,n are n independent observations on X, $\bar{\mathbf{X}}^t=(\bar{X}_{1i},\bar{X}_{2i},\ldots,\bar{X}_{pi})$ denote the sample mean matrix and $\boldsymbol{S}$ denotes the covariance matrix [14].

$$b_{1,p}=\frac{1}{n^2}\sum_{i,j=1}^{n}\left\{(\mathbf{X}_i-\bar{\mathbf{X}})^t\mathbf{S}^{-1}(\mathbf{X}_j-\bar{\mathbf{X}})\right\}^3 \quad (9)$$

Then in multivariate regression case, we express the multivariate skewness as in (10).

$$b_{1,2}=\frac{1}{n^2}\sum_{i,j=1}^{n}\left\{(\mathbf{e}_i-\bar{\mathbf{e}})^t\mathbf{S}^{-1}(\mathbf{e}_j-\bar{\mathbf{e}})\right\}^3 \quad (10)$$

where $\mathbf{e}_i^t=(\mathbf{e}_{1i},\mathbf{e}_{2i})$, i=1,2,…,148 are 148 residuals, $\bar{\mathbf{e}}^t=(\bar{\mathbf{e}}_1,\bar{\mathbf{e}}_2)$ denote the sample residuals mean matrix.

Meanwhile, the multivariate kurtosis is expressed in (11).

$$b_{2,p}=\frac{1}{n^2}\sum_{i=1}^{n}\left\{(\mathbf{X}_i-\bar{\mathbf{X}})^t\mathbf{S}^{-1}(\mathbf{X}_i-\bar{\mathbf{X}})\right\}^3 \quad (11)$$

Therefore, in multivariate regression case, we express the kurtosis as presented in (12).

$$b_{2,2}=\frac{1}{n^2}\sum_{i=1}^{n}\left\{(\mathbf{e}_i-\bar{\mathbf{e}})^t\mathbf{S}^{-1}(\mathbf{e}_i-\bar{\mathbf{e}})\right\}^3 \quad (12)$$

3.3. Statistics Software

Using equations showed in section 3.2, we use several software to help us in numerical calculations. To determine $\boldsymbol{S}$ or covariance matrix, we use Microsoft Excel. We obtained the following value of $\boldsymbol{S}$.

$$\mathbf{S}=\begin{bmatrix}2.45 & 0.13\\ 0.13 & 2.45\end{bmatrix}$$

Also, $b_{1,2}$ and $b_{2,2}$ can be determined using any algebra software. However, since having too many observations will be a little bit overwhelming, we need to choose several statistics software to help us in finding the value of $b_{1,2}$ and $b_{2,2}$. Cain, Zhang, and Yuan [15] provided us the macros for SAS, R and SPSS. For R, we can download the syntax and adjust it with our variables to get the output. For SPSS, we can download the macro, put it in in our device and then run it. We do the same steps for SAS.

p-ISSN 2798-253X
e-ISSN 2798-3153

Zhang, *et. al*. [16] also developed online calculator to determine the value of Mardia's skewness and kurtosis. We accessed the link of this online calculator freely and obtained the result directly. The link was retrieved from https://webpower.psychstat.org/models/kurtosis/. We chose our data file in any kinds of file types, and clicked the 'calculate' button. The result of the calculation through this calculator is presented in Fig.3.

```
Sample size:  148
Number of variables:  2

Univariate skewness and kurtosis
     Skewness   SE_skew  Kurtosis   SE_kurt
e1 -0.4860983 0.1993458 -0.060408 0.3961496
e2  1.5503263 0.1993458  3.700492 0.3961496

Mardia's multivariate skewness and kurtosis
                 b         z      p-value
Skewness  2.620028 64.627369 3.083089e-13
Kurtosis 11.528782  5.366186 8.041916e-08
```

Fig. 3. Skewness and Kurtosis value.

Based in Fig 3., the value of $b_{1,2}$ =2.620028 and $b_{2,2}$ =11.528782. Stated in Mardia's table in [17], we have these critical values; $b_{1,2,0.05,148}$=0.4, lower $b_{2,2,0.05,148}$=6.858, and upper $b_{2,2,0.05,148}$=9.3. For skewness, the sample is from multivariate normal distribution if the statistic value is less than critical value, while for kurtosis, the sample is from normal distribution if the statistic value is between lower critical value and upper critical value. Because the value of skewness is greater than 0.4 and kurtosis value is not in range [6.858, 9.3], residuals in our case do not follow multivariate normal distribution.

Besides using Mardia's table, we can also use the p-value taken from the result as in Fig.3. P-value of skewness and kurtosis are less than the significance level. This result indicates that we reject the null hypothesis. Based on the hypothesis stated in section 3.2, we can draw a conclusion that multivariate normality assumption for residuals in our case is not fulfilled.

4. Conclusion

Based on the results and discussion, multivariate skewness and kurtosis test are adaptive enough to be applied in multivariate regression. The more observations we have, the more difficult we determine the value of skewness and kurtosis because we have to calculate as many combinations of i and j matrix operations as possible, including subtraction, addition, inversion and multiplication. However, this difficulties are no longer a problem because we can use R, SPSS and SAS though macro and the online calculator. Using these tools, we are not required to do the comparison with another multivariate normality assumption test. Thus, it is expected that the future research of this comparison topic gain a better result.

References

[1] Field. A, Discovering Statistics Using SPSS, London: Sage Publication, 2009.

[2] K.V. Mardia, "Measures of multivariate skewness and kurtosis with applications", *Biometrika*, vol.57, no.3, pp. 519-530, Dec 1970.

[3] E.C. Alexopoulos, "Introduction to multivariate regression analysis", *Hippokratia*, vol.14, pp.23-28, December 2010.

[4] J. Osborne and E. Waters, "Four assumptions of multiple regression that researchers should always test", *Pract. Assess. Res. Eval*., vol. 8, Jan 2002.

[5] D.S. Gregory and H.M. Jackson, "Logistic and linear regression assumptions: violation recognition and control", *in Proc. Pharma. SAS Users Group (PharmaSUG)*, Jun. 16-19, Pennsylvania, 2019. p.21.

[6] K.R. Das and A.H.M.R. Imon, "A brief review of test for normality", *Am. J. Theor. Appl. Stat*., vol. 5, no. 1, pp. 5-12. Jan. 2016.

[7] Budiyono, *Statistika untuk Penelitian*, Surakarta: UNS Press, 2009.

[8] M.Williams, C.A.G. Grajales, and D. Kurkiewicz, "Assumptions of multiple regression: correcting two misconceptions", *Pract. Assess. Res. Eval*., vol. 18, 2013.

[9] H.Y. Kim, "Statistical notes for clinical researchers: assessing normal distribution (2) using skewness and kurtosis", *Restor. Dent. Endo*., vol. 38, no. 1, pp. 52-54, Feb. 2003.

[10] C.D. Lin, "Conducting test in multivariate regression", *in Proc. SAS Globar Forum*. April 28-May 1, Texas, p. 13.

[11] J.L. Romeu and A. Ozturk, "A comparative study of goodness-of-fit for multivariate normality", *J. Multivar. Annal.*, vol.46, pp. 309-334, Aug. 1993.

[12] U. Khasanah, "Penggunaan analisis regresi multivariat untuk memodelkan faktor-faktor yang memperngaruhi hasil belajar", B.S. thesis, Math. Educ., Universitas PGRI Semarang, Semarang, INA.

[13] A.C. Rencher, Multivariate Statistical Inference and Applications, Canada: John Wiley and Sons, Inc. 1998.

[14] K.V. Mardia, "Assessment of multinormality and robustness of Hotelling's T^2 test", *J. R. Stat. Soc. Series C (Appl. Stats.)*,vol. 24, no. 2, pp. 163-171, 1975.

[15] M.K. Cain, Z. Zhang, and K.H. Yuan, "Univariate and multivariate skewness and kurtosis for measuring normality: prevalence, influence and estimastion", *Behav. Res. Methods*, vol. 17, Oct 2016.

[16] Z.J. Zhang, K.H. Yuan, Y.Mai, M.K. Cain, H.Du, G.Jiang, H.Liu, A.Santoso, M.yang, X.Wang, and D. Mattew. "Univariate and Multivariate Skewness and Kurtosis Calculation". Web power: Analysis Online. Retrieved from https://webpower.psychstat.org/models/kurtosis/ on Feb 17, 2021.

[17] K.V. Mardia, "Application of some measure of multivariate skewness and kurtosis in testing normality and robustness studies", *Sankhya: Indian J. Stat. Series B*, vol. 36, no. 2, pp. 115-128, May 1974.

p-ISSN 2798-253X
e-ISSN 2798-3153