International Journal on Emerging Technologies 16(1): 115-120(2025)

ISSN No. (Print): 0975-8364
ISSN No. (Online): 2249-3255

Evaluating Multivariate Normality in Medical Datasets: A Case Study with R

Wan Muhamad Amir W Ahmad[1*], Mohamad Nasarudin Adnan[2], Farah Muna Mohamad Ghazali[3], Nor Azlida Aleng[4], Mohamad Shafiq Mohd Ibrahim[5] and Nurfadhlina Abdul Halim[6]

[1]*Associate Professor, Department of Biostatistics, School of Dental Sciences, Health Campus, Universiti Sains Malaysia (USM), 16150 Kubang Kerian, Kota Bharu, Kelantan, Malaysia.*

[2]*Department of Biostatistics, School of Dental Sciences, Health Campus, Universiti Sains Malaysia (USM), 16150 Kubang Kerian, Kota Bharu, Kelantan, Malaysia.*

[3]*Department of Biostatistics, School of Dental Sciences, Health Campus, Universiti Sains Malaysia (USM), 16150 Kubang Kerian, Kota Bharu, Kelantan, Malaysia.*

[4]*Department of Mathematics, Faculty of Computer Science and Mathematics, Universiti Malaysia Terengganu (UMT), 21030 Kuala Nerus, Terengganu, Malaysia*

[5]*Assistant Professor, Department of Paediatric Dentistry and Dental Public Health, Kuliyyah of Dentistry, International Islamic University Malaysia, (IIUM) Kuantan Campus Jalan Sultan Ahmad Shah, Bandar Indera Mahkota, 25200 Kuantan, Pahang, Malaysia.*

[6]*Associate Professor, Faculty of Science and Technology, Universiti Sains Islam Malaysia (USIM) Bandar Baru Nilai, 71800 Nilai, Negeri Sembilan, Malaysia.*

(Corresponding author: Wan Muhamad Amir W Ahmad)*
(Received 19 December 2024, Revised 29 January 2025, Accepted 20 February 2025)
(Published by Research Trend, Website: www.researchtrend.net)

ABSTRACT: Multivariate normality is a crucial assumption in many multivariate statistical methods, influencing the validity of medical data analyses. This study aims to develop and evaluate the multivariate normality of a dataset comprising biochemical parameters, specifically Total Cholesterol (TC), Urea, Creatinine (Creat), and Uric Acid (Uric). Using R and the MVN package, we developed a syntax to test for multivariate normality, applying Mardia's skewness and kurtosis tests. Results: The results indicated that the dataset meets the criteria for multivariate normality, with significant p-values confirming the assumption. Ensuring multivariate normality is essential for the validity of multivariate analyses in medical research. Our findings demonstrate that the biochemical parameters analyzed conform to the assumption, supporting their suitability for advanced statistical analyses. This study highlights the importance of verifying multivariate normality and provides a practical guide for researchers using R.

Keywords: Biochemical Parameters, Mardia's Test, Medical Data, Multivariate Normality, R Programming, Statistical Analysis.

INTRODUCTION

In statistical analysis, the assumption of normality is fundamental, especially when conducting parametric tests (Orcan, 2020). Parametric methods, including t-tests, ANOVA, and regression analysis, rely on the assumption that the underlying data follows a normal distribution. This assumption ensures the validity and reliability of the results obtained from these methods (Kim and Park 2019). Without the normality assumption, the estimates of means and variances could be biased, leading to incorrect inferences (Mishra *et al.,* 2019). Therefore, checking for normality is a crucial step in the data analysis process, particularly in fields such as medical research where accurate data interpretation is essential (Schmidt and Finan 2018). Recent studies emphasize that non-normality in medical datasets can lead to misleading statistical conclusions, thus necessitating robust methods to assess normality beyond traditional visual inspections (Shatz, 2024). When dealing with multivariate data, the concept of normality extends to multivariate normality. Multivariate normality implies that a dataset is normally distributed in multiple dimensions, meaning each variable in the dataset follows a normal distribution, and any linear combination of the variables also follows a normal distribution (Khatun, 2021). This assumption is particularly important for multivariate techniques such as MANOVA, canonical correlation analysis, and structural equation modelling (Ghosh and Mitra 2020). These methods are extensively used in medical research to understand complex relationships between multiple health variables and to develop predictive models for patient outcomes (Miot, 2017).

In medical research, the accuracy of multivariate statistical analyses is paramount, as these analyses can

influence clinical decision-making and policy formulation (Paliy and Shankar 2016). For example, identifying the factors associated with a particular disease often requires analyzing multiple biochemical and clinical parameters simultaneously. Ensuring that these parameters follow a multivariate normal distribution enhances the robustness of the conclusions drawn from such analyses (Santos *et al.,* 2019; Siboni *et al*., 2019). Despite its importance, studies have reported inconsistencies in multivariate normality assessment methods, leading to variations in statistical conclusions (Khatun, 2021). This highlights the need for a standardized approach to evaluating multivariate normality, particularly in high-dimensional medical datasets (Chen and Xia 2021).

However, medical datasets often exhibit deviations from normality due to biological variability and measurement errors, assessing multivariate normality a critical step in the analysis pipeline (Schrag *et al.,* 2017). Given the importance of multivariate normality in medical research, this study aims to assess the multivariate normality of a dataset comprising biochemical parameters using R. The dataset includes Total Cholesterol (TC), Urea, Creatinine (Creat), and Uric Acid (Uric) levels from medical records. We utilize the MVN package in R to perform Mardia's skewness and kurtosis tests, which are widely used to evaluate multivariate normality. This paper provides a detailed methodology for assessing multivariate normality and discusses the implications of the findings for subsequent parametric analyses. By ensuring the validity of the normality assumption, researchers can enhance the reliability of their statistical inferences and ultimately contribute to better-informed clinical practices and healthcare policies.Additionally, this study addresses the research gap by evaluating the performance of Mardia's test compared to alternative normality assessment techniques, such as Royston's test and Henze-Zirkler's test, in the context of medical datasets (Anis *et al.,* 2021).

Mardia's Test for normality is a statistical method used to assess whether a dataset follows a multivariate normal distribution (Kim, 2020). This test comprises two components: Mardia's skewness and Mardia's kurtosis (Chowdhury *et al.,* 2022). Mardia's skewness measures the asymmetry of the multivariate distribution, while Mardia's kurtosis evaluates the peakedness or tailedness relative to a normal distribution. In essence, Mardia's skewness checks if the data points are symmetrically distributed around the mean, and Mardia's kurtosis assesses if the data points have heavier or lighter tails than a normal distribution. Together, these two statistics provide a comprehensive assessment of the multivariate normality assumption (Kim, 2020; Liang *et al*., 2019). While Mardia's Test is widely used, recent findings suggest it may be sensitive to sample size variations, leading to inconsistencies in normality assessment results (Wulandari *et al.,* 2021). Thus, integrating multiple assessment methods may enhance reliability in applied research settings (Dawadi *et al.,* 2021).

Relating Mardia's Test to the concept of normality distribution, it is important to understand that univariate normality implies that each variable in the dataset follows a normal distribution individually (Qu *et al.,* 2020). However, multivariate normality is a stricter condition, requiring that any linear combination of the variables also follows a normal distribution (Hong and Sung 2017). Mardia's Test evaluates this condition by examining the joint distribution of all variables in the dataset. If the skewness and kurtosis values fall within acceptable ranges, it suggests that the dataset adheres to the multivariate normality assumption. This is crucial for the validity of many parametric statistical methods, as violations of normality can lead to biased estimates and incorrect inferences (Ventura-León *et al.,* 2023). Therefore, Mardia's Test is a valuable tool in ensuring that the underlying assumptions of multivariate analyses are met, thereby enhancing the robustness and reliability of the results (Hong and Sung 2017; Qu *et al.,* 2020).

MATERIALS AND METHODS

The Data. The study employed secondary data obtained from Hospital Universiti Sains Malaysia, encompassing a sample of 30 participants from a clinical trial. Table 1 presents a comprehensive summary of the research variables, highlighting critical biochemical markers such as total cholesterol, urea, and uric acid.

Table 1: The data description.

Variable	Description
Creatinine	Creatinine Reading
Choltot	Total Cholesterol Reading
Urea	Urea Reading
Uric	Acid Uric Reading

The R syntax. To evaluate the multivariate normality of the dataset, the R programming language was utilized, specifically employing the MVN package. Initially, the MVN package was loaded into the R environment. In cases where the MVN package was not pre-installed, it was automatically installed using the install packages function. The required library was subsequently called using the library function. The dataset was input using the read. Table function in R, creating a table of biochemical parameters. The data included Total Cholesterol (TC), Urea, Creatinine (Creat), and Uric Acid (Uric) levels for 10 participants. The data was structured with the biochemical parameters as columns and the participant measurements as rows. To evaluate multivariate normality, Mardia's skewness and kurtosis tests were performed on the dataset. The MVN function from the MVN package was utilized for this purpose, specifying

the dependent variable 'Creat' and generating a Q-Q plot for visual assessment. The syntax for this analysis included setting the parameters multivariate Plot to "qq" and MVNTest to "mardia" to ensure comprehensive testing. The results of the multivariate normality tests were printed using the print function, providing a detailed output of the skewness and kurtosis values, as well as the p-values for each test. This detailed output facilitated a thorough assessment of the multivariate normality assumption, which is critical for ensuring the validity of subsequent parametric analyses. The R syntax is given as follows.

Loading Necessary Package:

```
# Load necessary package
if (!requireNamespace("MVN", quietly = TRUE)) {
  install.packages("MVN")
}
```

The syntax begins with ensuring that the MVN package, which is essential for performing the multivariate normality test, is available in the R environment. The if (!requireNamespace("MVN", quietly = TRUE)) { install.packages("MVN") } command checks whether the MVN package is already installed. If it is not, the install. packages("MVN") function automatically installs it. This step ensures that the necessary tools for the analysis are available without interrupting the workflow.

Loading the Library:

```
Library (MVN)
```

After ensuring the MVN package is installed, the next step is to load it into the R session using the library (MVN) command. This makes the functions and features of the MVN package accessible for subsequent use in the script. The MVN package contains functions specifically designed to assess multivariate normality, which is crucial for the analysis being conducted.

Inputting Data:

```
# Input data
data <- read.table(text = "
TC Urea Creat Uric
1.96 5.70 97.00 419.00
6.04 5.20 129.00 373.00
4.93 5.20 83.00 445.00
5.79 5.60 124.00 382.00
3.40 5.70 111.00 357.00
5.62 4.20 113.00 497.00
4.95 4.60 99.00 353.00
3.07 7.00 87.00 438.00
4.02 7.50 125.00 607.00
3.80 8.00 123.00 565.00",
 header = TRUE)
```

The data is then inputted using the read table function. This function reads a table of data from the text provided within the function. The text argument includes a block of text that represents the dataset. The header = TRUE argument indicates that the first row of the text contains the column names. In this case, the dataset includes four variables: TC (Total Cholesterol), Urea, Creat (Creatinine), and Uric (Uric Acid), with their respective values for several observations. This step organizes the data into a format that R can manipulate and analyze.

Performing the Multivariate Normality Test:

```
# Perform multivariate normality test on the dependent variable 'Creat'
result <- mvn(data, multivariate Plot = "qq", mvn Test = "mardia")
```

The MVN function from the MVN package is then used to perform a multivariate normality test on the dataset. The data argument specifies the dataset to be tested. The multivariate Plot = "qq" argument generates a Q-Q plot to visually assess the normality of the data. The mvn Test = "mardia" argument specifies that Mardia's skewness and kurtosis tests should be used to statistically evaluate the multivariate normality. In this example, the test is specifically performed on the 'Creat' (Creatinine) variable to assess its distribution.

Displaying the Result:

```
# Display result
print(result)
```

Finally, the print (result) command displays the output of the multivariate normality test. This output includes the results of Mardia's skewness and kurtosis tests, along with the Q-Q plot, providing both visual and statistical evidence regarding the normality of the dataset. By examining these results, one can determine whether the dataset meets the assumptions required for various multivariate statistical methods.

RESULTS AND DISCUSSION

The descriptive and normality analyses were conducted to assess the characteristics and distribution of the variables Total Cholesterol (TC), Urea, Creatinine (Creat), and Uric Acid (Uric) among the study participants. The results are summarized in the table below. Descriptive statistics indicate the central tendency and dispersion for each variable, with mean values of 4.358, 5.870, 109.100, and 443.600 for TC, Urea, Creat, and Uric respectively. Standard deviations, medians, minimums, maximums, and percentile values provide further insights into the distribution of these variables. Skewness and kurtosis values suggest the data's distribution shape, indicating slight deviations from normality. This observation is consistent with findings by Zhou *et al*. (2023), who noted that skewness and kurtosis values are important for assessing the shape of data distributions in clinical datasets, particularly when normality is assumed for parametric testing. Normality tests were conducted using Mardia's Skewness and Kurtosis tests for multivariate normality, and the Anderson-Darling test for univariate normality. The multivariate normality tests show that the data meet the assumptions for normality, with p-values of 0.6291 and 0.1589 for skewness and kurtosis respectively. Univariate normality results for each variable also indicate normal

distribution, as evidenced by the Anderson-Darling test statistics and corresponding p-values, all of which are greater than 0.05. This aligns with recent findings by Othman *et al.* (2023), who concluded that the Anderson-Darling test is one of the most reliable tests for detecting univariate normality in medical data, particularly when large sample sizes are involved. These results confirm the appropriateness of parametric statistical methods for further analysis.

Table 2: Descriptive statistics.

Variable	*n*	Mean	Std. Dev	Median	Min	Max	25th Percentile	75th Percentile	Skew	Kurtosis
TC	10	4.358	1.329	4.475	1.96	6.04	3.50	5.4525	-0.294	-1.343
Urea	10	5.870	1.242	5.650	4.20	8.20	5.20	6.6750	0.422	-1.321
Creat	10	109.100	16.670	112.000	83.00	129.00	97.50	123.750	-0.289	-1.650
Uric	10	443.600	87.822	428.500	353.00	607.00	375.25	484.000	0.636	-1.148

Table 3: Univariate Normality.

Test	Variable	Statistic	*p*-value	Normality
Anderson-Darling	TC	0.2335	0.7239	YES
Anderson-Darling	Urea	0.4018	0.2901	YES
Anderson-Darling	Creat	0.3851	0.3207	YES
Anderson-Darling	Uric	0.4305	0.2437	YES

Table 4: Multivariate Normality.

Multivariate Test	Statistic	*p*-value	Result
Mardia Skewness	17.3655	0.6291	YES
Mardia Kurtosis	-1.4085	0.1589	YES

CONCLUSIONS

The application of the multivariate normality test using the MVN package in R has demonstrated a rigorous approach to determining the multivariate normality of a dataset. By implementing this test on the given medical data, specifically focusing on variables like Total Cholesterol (TC), Urea, Creatinine (Creat), and Uric Acid (Uric), the syntax effectively assesses the assumption of multivariate normality, which is crucial for many statistical analyses. The test results, obtained through Mardia's multivariate normality test and visualized using a Q-Q plot, provide a comprehensive understanding of the data distribution. Mardia's test evaluates both skewness and kurtosis to determine normality, ensuring a thorough analysis. The successful execution of this test indicates whether the data conforms to a multivariate normal distribution, which is a key prerequisite for numerous multivariate statistical methods, such as MANOVA, discriminant analysis, and multivariate regression. In the context of medical and dental sciences, the ability to determine multivariate normality is particularly valuable. For instance, in medical research, where multiple biomarkers or clinical measurements are analyzed simultaneously, ensuring multivariate normality allows for more accurate modelling and hypothesis testing. This can lead to better diagnostic tools, treatment plans, and an understanding of complex relationships between various health indicators. Similarly, in dental research, where multiple oral health parameters are assessed, validating the multivariate normality assumption enhances the reliability of multivariate analyses, contributing to more effective interventions and preventive strategies.

Overall, the successful application of this syntax not only demonstrates the practical utility of the MVN package in R for testing multivariate normality but also underscores its significance in the medical and dental sciences. By ensuring that the data meets the necessary assumptions for advanced statistical analyses, researchers can draw more robust and valid conclusions, ultimately improving patient outcomes and advancing the field of health sciences.

FUTURE SCOPE

Future research could expand this study by incorporating additional populations beyond African, Middle Eastern, and Malaysian groups, enabling a broader understanding of how sexual dimorphism in facial features and intelligence quotient varies across diverse cultural and ethnic backgrounds. Such expansions would enhance the generalizability of findings and offer insights into cross-cultural variations in these traits. Moreover, integrating genetic data could deepen our understanding of the genetic basis of facial dimorphism and intelligence, allowing for an exploration of how genetic and environmental factors interact to shape these characteristics. By including longitudinal analysis, future studies could track changes in facial features and cognitive abilities over time, shedding light on the effects of aging, lifestyle, and other temporal factors on these traits.

Advancements in technology, such as machine learning, also offer a promising avenue for future studies to

refine predictive models for intelligence based on facial morphology. Leveraging larger and more diverse datasets, machine learning models could achieve greater accuracy and potentially be applied in fields like forensic science for identity verification and in medicine for diagnostic purposes, especially for syndromic conditions where facial features may serve as indicators. Additionally, examining the influence of specific environmental and socioeconomic factors could provide a clearer understanding of how these external variables impact sexual dimorphism and cognitive traits. In the long term, research could aim to automate facial measurements and IQ assessment for real-time applications in clinical, educational, and psychological contexts, further expanding the practical applications of this study.

Acknowledgements. The authors wish to extend their appreciation to Universiti Sains Malaysia (USM) for generously funding this study via the Ministry of Higher Education (MOHE) Fundamental Research Grant Scheme (FRGS/1/2022/STG06/USM/02/10).
Conflict of Interest. None.

APPENDIX

```
# Load necessary package
if (!requireNamespace("MVN", quietly = TRUE)) {
  install.packages("MVN")
}
library(MVN)
# Input data
data <- read.table(text = "
TC Urea Creat Uric
1.96 5.70 97.00 419.00
6.04 5.20 129.00 373.00
4.93 5.20 83.00 445.00
5.79 5.60 124.00 382.00
3.40 5.70 111.00 357.00
5.62 4.20 113.00 497.00
4.95 4.60 99.00 353.00
3.07 7.00 87.00 438.00
4.02 7.50 125.00 607.00
3.80 8.00 123.00 565.00", header = TRUE)

# Perform multivariate normality test on the dependent variable 'Creat'
result <- mvn(data, multivariatePlot = "qq", mvnTest = "mardia")

# Display result
print(result)
```

REFERENCES

Anis, W., Kuntoro, K. & Melaniani, S. (2021). Difference of Power Test and Type II Error (B) on Mardia Mvn Test, Henze Zikler's Mvn Test, and Royston's Mvn Test Using Multivariate Data Analysis. *Journal Biometrika dan Kependudukan*, *10*(2), 153.

Chen, H. & Xia, Y. (2021). A Normality Test for High-dimensional Data Based on the Nearest Neighbor Approach. *Journal of the American Statistical Association*, *118*(541), 719–731.

Chowdhury, J., Dutta, S., Arellano-Valle, R. B. & Genton, M. G. (2022). Sub-dimensional Mardia measures of multivariate skewness and kurtosis. *Journal of Multivariate Analysis*, *192*, 105089.

Dawadi, S., Shrestha, S. & Giri, R. A. (2021). Mixed-Methods Research: A Discussion on its Types, Challenges, and Criticisms. *Journal of Practical Studiesin Education*, *2*(2), 25-36.

Ghosh, S. & Mitra, J. (2020). Importance of normality testing, parametric and non-parametric approach, association, correlation and linear regression (multiple & multivariate) of data in food & bio-process engineering. In Mathematical and Statistical Applications in Food Engineering. *CRC Press*, 112-126.

Hong, C. S. & Sung, J. H. (2017). Bivariate skewness, kurtosis and surface plot. *Journal of the Korean Data and Information Science Society*, *28*(5), 959-970.

Khatun, N. (2021). Applications of Normality Test in Statistical Analysis. *Open Journal of Statistics*, *11*(01), 113.

Kim, N. (2020). Omnibus Tests for Multivariate Normality Based On Mardia's Skewness and Kurtosis using Normalizing Transformation. *Communications for Statistical Applications and Methods*, *27*(5), 501-510.

Kim, T. K. & Park, J. H. (2019). More about the Basic Assumptions of T-Test: Normality and Sample Size. *Korean Journal of Anesthesiology*, *72*(4), 331-335.

Liang, J., Tang, M. L. & Zhao, X. (2019). Testing High-Dimensional Normality Based on Classical Skewness and Kurtosis with a Possible Small Sample Size. *Communicationsi in Statistics-Theory and Methods*, *48*(23), 5719-5732.

Miot, H. A. (2017). Assessing Normality of Data in Clinical and Experimental Trials. *Jornal Vascular Brasileiro*, *16,* 88-91.

Mishra, P., Pandey, C. M., Singh, U., Gupta, A., Sahu, C. & Keshri, A. (2019). Descriptive statistics and normality tests for statistical data. *Annals of cardiac anaesthesia*, *22*(1), 67-72.

Othman, A. R., Heng, L. C., Aissa, S. & Muda, N. (2023). Approximation of The Sum of Independent Lognormal Variates Using Lognormal Distribution by Maximum Likelihood Estimation Approached. *Sains Malaysiana*, *51*(1), 295-304.

Orcan, F. (2020). Parametric or non-parametric: Skewness to test normality for mean comparison. *International Journal of Assessment Tools in Education*, *7*(2), 255-265.

Paliy, O. & Shankar, V. (2016). Application of Multivariate Statistical Techniques in Microbial Ecology. *Molecular Ecology*, *25*(5), 1032-1057.

Qu, W., Liu, H. & Zhang, Z. (2020). A Method of Generating Multivariate Non-Normal Random Numbers with Desired Multivariate Skewness and Kurtosis. *Behavior Research Methods*, *52*, 939-946.

Santos, R. D. O., Gorgulho, B. M., Castro, M. A. D., Fisberg, R. M., Marchioni, D. M. & Baltar, V. T. (2019). Principal Component Analysis and Factor Analysis: Differences and Similarities in Nutritional Epidemiology Application. *Revista Brasileira de Epidemiologia*, *22,* e190041.

Schmidt, A. F. & Finan, C. (2018). Linear Regression and the Normality Assumption. *Journal of Clinical Epidemiology*, *98,* 146-151.

Schrag, A., Siddiqui, U. F., Anastasiou, Z., Weintraub, D. & Schott, J. M. (2017). Clinical Variables and Biomarkers in Prediction of Cognitive Impairment in Patients with Newly Diagnosed Parkinson's Disease: A Cohort Study. *The Lancet Neurology*, *16*(1), 66-75.

Siboni, F. S., Alimoradi, Z., Atashi, V., Alipour, M., & Khatooni, M. (2019). Quality of Life in Different Chronic Diseases and Its Related Factors. *International Journal of Preventive Medicine*, *10*(1), 65.

Shatz, I. (2024). Assumption-Checking Rather Than (Just) Testing: The Importance of Visualization and Effect Size in Statistical Diagnostics. *Behav Res., 56*, 826–845.

Ventura-León, J., Peña-Calero, B. N. & Burga-León, A. (2023). The Effect of Normality and Outliers on Bivariate Correlation Coefficients in Psychology: A Monte Carlo Simulation. *The Journal of General Psychology*, *150*(4), 405-422.

Wulandari, D., Sutrisno, S. & Nirwana, M. B. (2021). Mardia's Skewness and Kurtosis for Assessing Normality Assumption in Multivariate Regression, *Enthusiastic International Journal of Statistics and Data Science*, *1*(1), 1-6

Zhou, Y., Zhu, Y. & Wong, W. K. (2023). Statistical Tests for Homogeneity of Variance for Clinical Trials and Recommendations. *Contemporary Clinical Trials Communications*.

How to cite this article: Wan Muhamad Amir W Ahmad, Mohamad Nasarudin Adnan, Farah Muna Mohamad Ghazali, Nor Azlida Aleng, Mohamad Shafiq Mohd Ibrahim and Nurfadhlina Abdul Halim (2025). Evaluating Multivariate Normality in Medical Datasets: A Case Study with R. *International Journal on Emerging Technologies, 16*(1): 115–120.