International Journal of Scientific Research in Science, Engineering and Technology
Print ISSN - 2395-1990
Online ISSN : 2394-4099
Available Online at : www.ijsrset.com
doi : https://doi.org/10.32628/IJSRSET2411213

Analysis of Multivariate Data Variance Detected Outlier to Determine Factors Influencing Lipid Profile in Diabetes Patients

Rahmah Humaeroh[1], Muhammad Nur Aidi[2], Bagus Sartono[3], Nazarina[4]
[1,2,3]Department of Statistics, IPB university, Bogor, Indonesia
[4]National Research and Inovation Agency, Bogor, Indonesia
Email : muhammadai@apps.ipb.ac.id

ARTICLEINFO

Article History:

Accepted: 10 March 2024
Published: 30 March 2024

Publication Issue :
Volume 11, Issue 2
March-April-2024
Page Number :
147-154

ABSTRACT

Diabetes is characterized by high blood glucose levels and can lead to cardiovascular complications. This research, will be investigating he factors influencing lipid profile in diabetes patients and the differences in these profiles under various treatments. The research used the MANOVA test to analyze differences in lipid profile under different treatment and controlled for covariates using Mancova. They also used robust methods to address outliers is minimum covariance determinant. The result suggest that gender is a significant factor influencing lipid profile in diabetes patients. The most effective analytical methods were found to be robust manova and robust mancova, with low RMSE values indicating their accuracy.
Keywords: Diabetes, Outlier, MANOVA, MANCOVA

I. INTRODUCTION

Diabetes is a chronic disease characterized by hyperglycemia, in which blood glucose levels exceed normal limits due to insulin deficiency, insulin resistance, or both [1]. Diabetes must require lifelong management to prevent various chronic complications, such as disorders in both macrovascular and microvascular blood vessels [2] [3].

The research conducted by Aryal states that an increase in the ratio of triglycerides, HDL, and LDL in individuals with diabetes may indicate a higher risk of cardiovascular disease [4]. Therefore, it is essential to identify the factors impacting the lipid profile in individuals with diabetes to prevent it. When response variables exceed one or more predictor variables, the method used is MANOVA, and when covariates are included, the method used is MANCOVA. In general, observational data often contain outliers, and one of the methods used is the minimum covariance determinant estimator.

Ningrum (2009) conducted a study using the Minimum Covariance Determinant Estimator (MCD) method in multivariate analysis, particularly in robust canonical correlation analysis [5]. This study refers to the research by Todorov and Filzmoser (2010) titled

"Robust statistic for the one-way MANOVA". In that study, two methods were used to estimate Wilk's Lambda test statistic to handle outlier data in MANOVA: rank transformation and Minimum Covariance Determinant Estimator (MCD). The results of the analysis indicated that the MCD estimator produced better tests.

In this study, a test will be conducted on data from individuals with diabetes using MANOVA and MANCOVA methods on data that meets the assumption of normal multivariate distribution and detects outliers using robust methods. The aim of this research is to determine the factors that influence the levels of High Density Lipoprotein (HDL), Low Density Lipoprotein (LDL), and Triglycerides (TG) in individuals with diabetes with detected outlier data and the best method based on the lowest RMSE value.

II. METHODS AND MATERIAL

This study uses secondary data on people with diabetes taken from 5 villages in Central Bogor District. With research variable :

Table 1. Variables used in the study

Variable	Variable Name
Y_1	HDL
Y_2	LDL
Y_3	Triglycerides
X_1	Sleep Disorders
X_2	Consumption of diabetes drugs
X_3	Smoking Status
X_4	Physical Activity
X_5	Sport
X_6	Gender
X_7	Duration of diabetes
X_8	Fat Intake
X_9	Carbohydrate Intake
X_{10}	Vegetable Fruit Intake
X_{11}	Blood pressure
X_{12}	Feeding behavior

Manova

Manova is a generalized from of anova that measures more than one response variable in each experimental unit [6]. Manova examines the effect of treatment applied to more than one response variable by considering the dependence between response variables [7]. The model is

$$\boldsymbol{y_{ij} = \mu + \tau_i + \varepsilon_{ij}}$$

With y_{ij} reprents the observation for dependent variable I from group j,μ denotes the overall mean, τ_I is the effect of the i treatment, ε_{ij} is the error term.

Mancova

Mancova is a statistical technique that is utilized to examine the variations in treatment effect that have been adjusted for covariates. The incorporation of covariates in the model has the advantage of decreasing the sources of error in the covariance matrix, which ultimately enhances the precision of the design [8]. The Mancova model is a

$$\boldsymbol{y_{ij} = \mu + \tau_i + Bx_{ij} + \varepsilon_{ij}}$$

With y_{ij} is represents the observation for dependent variable i from group j, μ denotes the overall mean, τ_I is the effect of the i treatment, Bx_{ij} represents the effect of the covariate and ε_{ij} stands for the random error.

Minimum Covariance Determinant Estimator

The Minimum Covariance Determinant (MCD) method in multivariate analysis is used to find robust estimates of the center and spread of multivariate data, especially in situations where data may affected by outliers [9]. The MCD estimator is generated from the FAST-MCD algorithm. MCD is the pair of **t(x)** and **C(X)** of an observational h-sized subsample with the smallest variation-demonstration matrix determinant. The limit of the sub-sample hose h is with

$$h_0 \leq h \leq n\, h_0 = ((n+p+1)/2)$$

With

$$T_1 = \frac{1}{h} \Sigma_{i \in H1} y_i$$

$$S_1 = \frac{1}{h} \Sigma_{i \in H1} (y_i - T_1)(y_i - T_1)'$$

The MCD estimator looks for the subset of x number h of the element where h is the smallest integer of $((n+p+1)/2)$.

RMSE

The accuracy of the model Manova,Manova robutst, Mancova and Mancova robust methods is measured using RMSE with the formula :

$$\sqrt{\sum_{i=1}^{n} \frac{(\hat{Z}(x_0) - Z(x_i))^2}{n}}$$

With $\hat{Z}(x_0)$ as the estimated predicted val $Z(x_i)$ as the observed value at point i,and n as the number of sample used [10].

Analysis Procedure

1. Preparation/Pre-analysis
   Data cleaning *will be carried out* at this stage, such as handling *missing values*, checking *outliers* and others.
2. Describe the data
   Once the data is confirmed to be ready for processing, the next step is to describe the data for each variable
3. Perform assumption testing
4. Perform outlier *detection testing*
5. Mancova analysis with *Minimum Covariance Determinant (MCD) estimator*
6. Testing hypotheses with *Wilk's Lambda test statistics with* the Mancova method with classical estimators of the covariance matrix and MCD.
7. Compare RMSE values for manova, manova robust, mancova and mancova robust methods for the best method.

III.RESULTS AND DISCUSSION

The data used in this research comprises various characteristics for each variable obtained from the respondents. The information gathered pertains to sleep disorders, diabetes medication consumption, smoking habits, physical activity levels, sports participation, gender, and eating behavior. The table below presents the calculated percentages for each category of each variable.

Table 2. Characteristic variables

Variables	Percentage (%)
Sleep Disorders	
Yes	36,75
Not	63,25
Diabetes Drug Consump	
Yes	30,12
Not	69,88
Smoking Status	
Never	51,81
No, sometimes	14,46
No, it used to be ever	10,24
Sometimes	6,63
Every Day	16,87
Physical Activity	
Tall	60,24
Low	39,76
Sport	
Yes	42,17
Not	57,83
Gender	
Man	28,31
Woman	71,69
Eating behavior	
Good	95,78
Bad	4,22

The assumption checks

The result of the correlation test is demonstrated the relationship between the response variables which indicate that theres is a dependency between them. The small p-value with alpha 0.05 adds to the validity of the results. The correlations that exist are diverse, as shown by the correlogram graph in Figure 1.

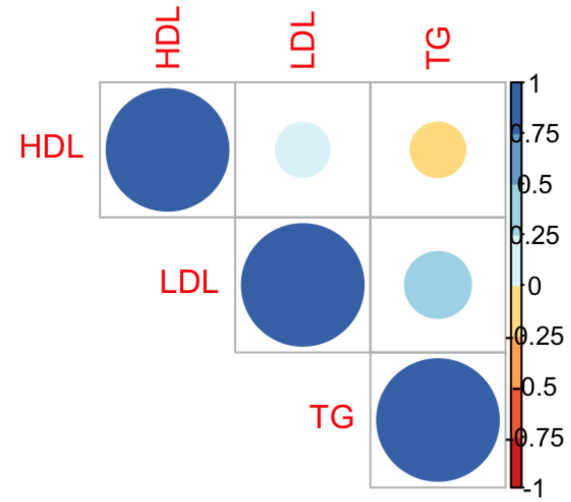

Figure1. Lipid Profile Correlogram

Based on the figure 1 the correlation values between the response variables, which indicate a weak positive correlation between HDL and LDL levels, as well as between HDL and TG levels. Furthermore, a positive correlation between LDL and TG levels is also discernible.

Multivariate Normality Assumption

Assessing the multivariate normality assumption is an essential aspect of multivariate statistical analysis. It is crucial to evaluate the distribution of the data to be analyzed. The obtained value of 0.0913 in the normality test suggests that the data follows a normal distribution.

Homogeneity of Variance-Covariance Matrices Assumption

Box's M test evaluates the homogeneity assumption of variance-covariance matrices. This analysis is conducted contingent upon the variables fulfilling the assumption of multivariate normality and expecting homogeneity of variance-covariance matrices across each treatment category. The outcome of the Box's M test indicates that the variance-covariance matrice.

Multivariate outlier detection

Outlier detection in multivariate analysis can be determined through the use of Mahalanobis distance values in comparison to $\chi_{p,}$ values. Furthurmore, this can also be visually represented on a graph, where points that fall outside of the boundaries are identified as outliers. The following are the results of Mahalanobis distance using both classical and robust MCD methods.

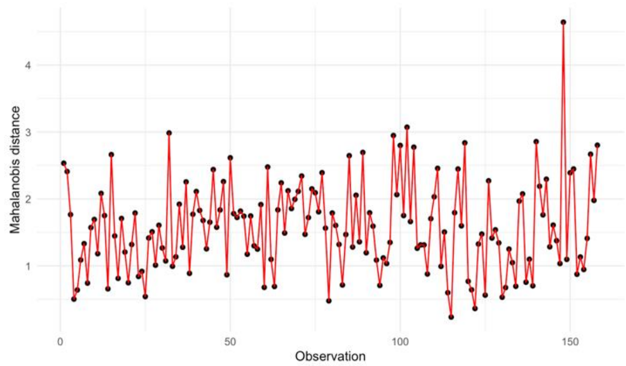

Figure 2. Mahalanobis distance

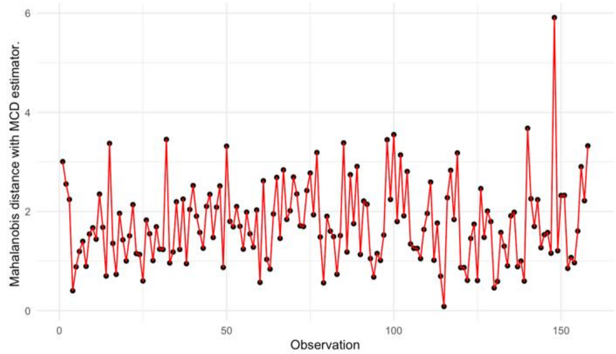

Figure 3. Mahalanobis distance with MCD estimator

In light of the figure 2 and 3 presented above, it is apparent that a number of data points are suspected of being outliers. Therefore, it is crucial to definitively identify which observations are deemed to be outliers.

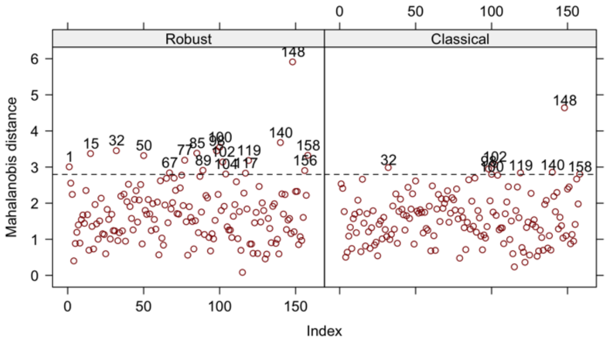

Figure 4. Outlier Detection

Based on the findings of the outlier detection analysis carried out using the MCD and classical approaches shown in Figure 4, points that lie above the line are considered outliers.

Manova

After testing the assumptions, the next step is MANOVA testing for the effect of treatment on people with diabetes using the Wilks Lambda statistic. The following is a table of MANOVA test results in people with diabetes.

Table 3. Manova

Effect	Wilks' Lambda	
	Value	p-value
Sleep Disorders	0,9818	0,4172
Diabetes Drug Consumt	0,9568	0,0779
Smoking Status	0,8480	0,0138
Physical Activity	0,9700	0,1955
Sport	0,9719	0,2214
Gender	0,8560	0,0000
Eating behavior	0,9937	0,8057

The Wilks Lambda test statistics in Table produce p-value = 0.0138 for smoking status and p-value = 0.0000 for gender in people with diabetes with alpha 5%. So, smoking status and gender have differences in lipid profiles in diabetes patients.

In the analysis conducted, an ANOVA test was performed to assess the average discrepancies in HDL, LDL, and TG levels among individuals with diabetes based on their smoking status and gender.

Table 4. Anova

Predictor Variable	Respons Variable	Db	F	P - value
	HDL	1	22,37	0,000
Smoking Status	LDL	1	0,4	0,528
	TG	1	0,459	0,499
	HDL	1	26.23	0,000
Gender	LDL	1	0,452	0,502
	TG	1	1,13	0,289

Based on the table above, the ANOVA test shows that smoking and gender have a significant effect on HDL levels.

Manova Robust

Manova Robust can be used when multivariate data has outliers, using MCD estimator method.

Table 4. Manova robust

Effect	Wilk's Lambda	
	Value	p-value
Sleep Disorders	0,9402	0,0744
Diabetes Drug Consumtion	0,9544	0,1491
Smoking Status	0,8269	0,0539
Physical Activity	0,9823	0,5521
Sport	0,9600	0,1923
Gender	0,8635	0,0008
Eating behavior	0,9415	0,0682

Based on Table 4, it can be observed that the Wilks Lambda statistic generated for gender is 0.8635. Meanwhile, the p-value for gender is 0.0008. Therefore, it can be concluded that gender influences the lipid profile of diabetes patients. Significant differences in gender on lipid profile were found in HDL levels, as shown in the following table.

Table 5. Anova robust

Predictor Variable	Respons Variable	P-value
	HDL	0,0000
Gender	LDL	0,4291
	TG	0,4658

Mancova

The application of covariates in the analysis of variance aims to eliminate indirect influences on the response variable that cannot be controlled and are solely direct effects of the experimental factors, thus providing more precise results. The MANCOVA test is essentially a MANOVA test with the inclusion of covariate effects.

Table 6. Mancova

Effect	Wilk's Lambda	
	Value	p-value
Sleep Disorders	0,972	0,264
Diabetes Drug Consumtion	0,963	0,151
Smoking Status	0,925	0,527
Physical Activity	0,992	0,768
Sport	0,952	0,073
Gender	0,913	0,005
Eating Behavior	0,996	0,894
Diabetes Duration	0,993	0,792
Fat Intake	0,982	0,477
Carbohydrate Intake	0,976	0,338
Vegetable Fruit Intake	0,992	0,779
Blood pressure	0,964	0,162

The Wilks' Lambda test statistics in Table 6 yielded a p-value of 0.005 for gender. That is, there was an effect of gender differences on blood lipid profiles in people with diabetes after conditioning the covariate to a fundamental level of 5%. Table 7 shows the ANCOVA test results as a follow-up analysis in determining differences in blood lipid prophy between gender.

Table 7. Ancova

Predictor Variable	Response Varariable	Db	F	P-value
	HDL	1	13,175	0,000
Gender	LDL	1	0,684	0,410
	TG	1	0,654	0,420

The test statistics in Table 12 produce p-value = 0.000 for gender factors against HDL levels at a fundamental level of 5%, meaning there is a difference in average HDL levels between the gender after controlling for covariates.

Mancova robust

Mancova robust can be used when multivariate data contains outlier. By using a robust approach, Mancova becomes more resistant to disturbances from outlier, thus providing more realible and stable estimates.

The Wilks' Lambda test statistics in Table 8 yielded a p-value of 0.011 for sex. That is, there was an effect of sex differences on blood lipid profiles in people with diabetes after conditioning the covariate to a fundamental level of 5%. Table 9 shows the ANCOVA test results as a follow-up analysis in determining differences in blood lipid prophy between gender.

Table 8. Mancova Robust test

Effect	Wilk's Lambda	
	Value	p-value
Sleep Disorders	0,970	0,229
Diabetes Drug Consumption	0,967	0,189
Smoking Status	0,943	0,755
Physical Activity	0,984	0,514
Sport	0,953	0,079
Gender	0,924	0,011
Eating behavior	0,992	0,785
Diabetes Duration	0,986	0,585
Fat Intake	0,987	0,594

Carbohydrate Intake	0,974	0,302
VegetableFruit Intake	0,990	0,695
Blood pressure	0,959	0,115

The test statistics in Table 9 produce p-value = 0.001 for sex factors against HDL at a real level of 5%, meaning there is a difference in average HDL levels between the sexes after controlling for covariate influence.

Table 9 Ancova Robust

Predictor Variables	Respons Variable	db	F	P-value
Gender	HDL	1	10,886	0,001
	LDL	1	0,397	0,530
	TG	1	0,051	0,822

RMSE

The best model has a smaller RMSE value. The RMSE values for each method are as follows :

Table 10. The RMSE values

	HDL	LDL	TG
Manova	0,301	1,630	5,357
Manova Robust	0,014	0,002	0,001
Mancova	0,300	1,654	5,345
Mancova Robust	0,014	0,002	0,001

From the RMSE values above, the best method is Mancova Robust because it has the lowest RMSE values for all variables (HDL, LDL, and TG), which are 0.014 for HDL, 0.002 for LDL, and 0.001 for TG. The lower the RMSE value, the better the model performance in predicting the true values in the dataset. Therefore, Mancova Robust can be considered as the best method in this case.

IV.CONCLUSION

MANOVA (Multivariate Analysis of Variance) indicates that smoking status and gender have significant differences in lipid profiles in diabetes patients. The Wilks' Lambda test statistics show significant p-values for smoking status (p-value = 0.0138) and gender (p-value = 0.0000).

ANOVA (Analysis of Variance) tests further reveal that smoking status and gender have significant effects on HDL levels in diabetes patients. The p-values for smoking status and gender are both 0.000. MANCOVA (Multivariate Analysis of Covariance) also demonstrates that gender influences the lipid profile of diabetes patients. The Wilks' Lambda test statistics yield a significant p-value of 0.005 for gender. Robust MANOVA and Robust MANCOVA can be used when multivariate data contains outliers. The Wilks' Lambda test statistics for gender in both robust methods yield significant p-values, indicating that gender differences affect blood lipid profiles in diabetes patients.

Based on the RMSE values, Mancova Robust is the best method as it has the lowest RMSE values for all variables (HDL, LDL, and TG). In summary, smoking status and gender significantly affect lipid profiles in diabetes patients, with gender having a powerful influence. Additionally, Mancova Robust is the best data analysis method due to its robustness and superior predictive performance.

V. REFERENCES

[1] Senja Atika Sari, Luthfiatil Fitri N, Ludiana L, Kesuma Dewi T, Immawati I. 2023. Edukasi Untuk Meningkatkan Pengetahuan Kader Kesehatan Tentang Diabetes Mellitus Dan Senam Kaki Diabetes. *Jurnal Masyarakat Madani Indonesia*. 2(3):135–138.Doi:10.59025/Js.V2i3.89.

[2] Piero Mn. 2015. Diabetes Mellitus – A Devastating Metabolic Disorder. *Asian Journal Of Biomedical And Pharmaceutical Sciences*. 4(40):1–7.Doi:10.15272/Ajbps.V4i40.645.

[3] Kharroubi At. 2015. Diabetes Mellitus: The Epidemic Of The Century. *World J Diabetes*. 6(6):850.Doi:10.4239/Wjd.V6.I6.850.

[4] Aryal M. 2010. Evaluation Of Non-Hdl-C And Total Cholesterol: Hdl-C Ratio As Cumulative Marker Of Cardiovascular Risk In Diabetes Mellitus. Volume Ke-9.
[5] Ningrum D. 2009. Analisis Korelasi Kanonik Robust Menggunakan Matriks Kovarian dengan Penduga Minimum Covariance Determinant (MCD) pada Data Pencilan. Malang.
[6] Mattjik Aa, Sumertajaya Im. 2011. *Sidik Peubah Ganda Dengan Menggunakan Sas*. Pertama. Wibawa Gna, Hadi Af, Editor. Bogor: Ipb Press.
[7] Aidi Muhammad Nur, Sumertajaya I Made, Wijayanto Hari. 2023. *Analisis Peubah Ganda Terapan Inferensi Peubah Ganda* . Ed Ke-1 Afyandi, Nurdiansyah M A, Alwedy M, Editor. Bogor: Ipb Press
[8] Rencher Ac. 2002. *Methods Of Multivariate Analysis*. 2nd Ed. New York : John Wiley & Sons,Inc.
[9] Rousseeuw, Van Driessen. 1999. 1999_Fastmcd_Technometrics. 41.
[10] Montgomery D, Peck Elizabeth A, Vining G Geoffrey. 2012. *Introduction To Linear Regression Analysis*. Fifth. Wiley.