We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

6,300

Open access books available

172,000

International authors and editors

190M

Downloads

Our authors are among the

154

Countries delivered to

TOP 1%

most cited scientists

12.2%

Contributors from top 500 universities

WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index
in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com

Computational Statistics with Dummy Variables

Adji Achmad Rinaldo Fernandes, Solimun and Nurjannah

Abstract

Cluster analysis is a technique commonly used to group objects and then further analysis is carried out to obtain a model, named cluster integration. This process can be continued with various analyzes, including path analyzes, discriminant analyzes, logistics, etc. In this chapter, the author discusses the reason to use dummy variables in this type of cluster analysis. Dummy variables are the main way that categorical variables are included as predictors in modeling. With statistical models such as linear regression, one of the dummy variables needs to be excluded, otherwise the predictor variables are perfectly correlated. Thus, usually if a categorical variable can take k values, we only need $k-1$ dummy variables, the k -th variable being redundant, it does not bring any new information. When more dummy variables than needed are used this is known as dummy variable trapping. The advantage to use dummy variables is that they are simple to use and the decision making process is easier to manage. The novelty in this chapter is the perspective of the dummy variable technique using cluster analysis in statistical modeling. The data used in this study is an assessment of the provision of credit risk at a bank in Indonesia. All analyzes were carried out using software R.

Keywords: dummy, cluster, integrated cluster with logistic regression, integrated cluster with discriminant analysis, integrated cluster with path analysis

1. Introduction

The application of cluster analysis is commonly used to group objects. Cluster analysis can be used to group objects and then further analysis is carried out to obtain a model, namely cluster integration. Cluster integration can be continued with various analyzes, including path analysis, discriminant analysis, logistics, etc. In cluster integration with path analysis, it aims to group homogeneous objects into one group, the goal is that the resulting residual variance is homogeneous in addition to maximizing the adjusted R^2 value. In cluster integration with discriminant analysis, the benefits of cluster analysis generated can maximize the accuracy, sensitivity, and specificity of the model. In this chapter, we will explain the technical perspective of dummy variables using cluster analysis in statistical modeling, such as regression analysis, path analysis, and discriminant analysis.

2. Why use dummy variables

Dummy variables are numerical variables that represent categorical data, such as gender, race, political affiliation, etc. Technically, the dummy variable is dichotomous, a quantitative variable. Their value range is small; they can only take two quantitative values. As a practical matter, regression results are easiest to interpret when the dummy variable is constrained to two specific values, 1 or 0. Typically, 1 represents the presence of a qualitative attribute, and 0 represents its absence. Categorical variables have more than two categories, can be represented by a set of dummy variables, with one variable for each category. Numerical variables can also be coded to explore nonlinear effects. Dummy variables are also known as indicator variables, design variables, contrasts, one-hot coding, and binary basis variables [1].

Dummy variables are the main way that categorical variables are included as predictors in modeling. For example, in linear regression analysis, the response variable is profit, and the predictor variable is employee group. With statistical models such as linear regression, one of the dummy variables needs to be excluded (by convention, the former, or the latter), otherwise, the predictor variables are perfectly correlated [2].

When defining dummy variables, a common mistake is to define too many variables. If a categorical variable can take on k values, then you tend to define k dummy variables. You only need $k-1$ dummy variable.

The k -th dummy variable is redundant; it does not bring any new information. And that creates a severe multicollinearity problem for analysis. Using k dummy variables when only $k-1$ dummy variables are needed is known as dummy variable trapping.

Regression analysis treats all independent variables (X) in the analysis as numerical. A numeric variable is an interval or ratio scale variable whose values can be directly compared, e.g. “10 is double 5,” or “3 minus 1 equals 2.” However, you may want to include a nominal scale attribute or variable such as: “Product Brand” or “Defect Type” in your study. Say you have three types of defects, numbered “1,” “2” and “3.” In this case, “3 minus 1” means nothing. You cannot subtract handicap 1 from handicap 3. The numbers here are used to indicate or identify the degree of “Type of Disability” and have no intrinsic meaning of their own. A dummy variable is created in this situation to “trick” the regression algorithm into the correct attribution of the analysis variable [3].

The main benefit of dummy variables is that they are simple. Often there are better alternative basis functions, such as orthogonal polynomials, effect coding, and splines. If dummy variables are used in linear regression analysis, then there are several advantages [4], including:

- a. The dependent variable prediction process becomes more focused and accurate, different from ordinary multiple regression
- b. Because the data is not qualitative, the prediction results are easy to interpret
- c. The decision-making process tends to be easy

3. Hierarchical cluster

Cluster analysis (group analysis) is an analytical method that aims to group objects into several groups, objects in groups are homogeneous (same) while other

group members are heterogeneous (different) [5]. The procedure for group formation in Cluster analysis is divided into two, namely hierarchical and non-hierarchical methods. Grouping with the hierarchical method is used when there is no information about the number of clusters. The main principle of the hierarchical method is to group objects that have something in common with one group. While the non-hierarchical method is used when information about the number of clusters is known or has been determined [6].

This method starts grouping with two or more objects that have the closest object. Then the process is continued by passing to another object that has second proximity. And so on to form a tree in which there is a hierarchy or level from the most similar to the different. The tree formed by this cluster is also called a dendrogram. This tree is useful for providing deeper clarity on the clustering process.

The stages of grouping data using the hierarchical method are [7]:

1. Determine k as the number of clusters to be formed.
2. Each object data is considered as a cluster so that $n = N$.
3. Calculate the distance between clusters.
4. Find two clusters that have the least distance between clusters and combine them (meaning $N = n-1$).
5. If $n > k$, then go back to step 3.

According to [6] in the method of forming groups in the hierarchical method, there are two approaches, namely agglomerative hierarchical methods (Agglomerative Hierarchical Methods) and divisive hierarchical methods (Divisive Hierarchical Methods). The agglomerative method starts by assuming that each object is a cluster. Then the two objects that have the closest distance are combined into one cluster. The process continues so that in the end it will form a cluster consisting of all objects.

4. Integrated cluster with logistic regression

4.1 Integrated cluster equation model with logistic regression analysis

The model of the integration of cluster analysis with logistic analysis of the dummy variable approach is the same as the general model of multiple linear regression analysis with dummy variables.

The general model of the integrated cluster with logistic analysis can be written in the following Eq. (1).

$$y_i = \frac{\exp \left(\begin{matrix} \beta_0 + \beta_1 x_{1i} + \dots + \beta_p x_{pi} + \\ D_1 \beta_{p+1} + D_1 \beta_{p+1} x_{1i} + \dots + D_1 \beta_{2p+1} x_{pi} + \\ D_2 \beta_{p+2} + D_2 \beta_{p+3} x_{1i} + \dots + D_2 \beta_{3p+2} x_{pi} + \\ \dots + D_q \beta_{pq+1} + D_q \beta_{pq+2} x_{1i} + \dots + D_q \beta_{p(q+1)} x_{pi} \end{matrix} \right)}{1 + \exp \left(\begin{matrix} \beta_0 + \beta_1 x_{1i} + \dots + \beta_p x_{pi} + \\ D_1 \beta_{p+1} + D_1 \beta_{p+1} x_{1i} + \dots + D_1 \beta_{2p+1} x_{pi} + \\ D_2 \beta_{p+2} + D_2 \beta_{p+3} x_{1i} + \dots + D_2 \beta_{3p+2} x_{pi} + \\ \dots + D_q \beta_{pq+1} + D_q \beta_{pq+2} x_{1i} + \dots + D_q \beta_{p(q+1)} x_{pi} \end{matrix} \right)} \quad (1)$$

where,
 y_i : response variable at the i -th observation unit
 x_{ki} : the k -th predictor variable on the i -th observation unit
 β_p : coefficient of the p -th logistic function
 D_q : q -th dummy variable
 p : number of predictor variables
 q : the number of clusters formed is reduced by 1
 i : 1, 2, 3, ..., n

4.2 Logistics regression analysis assumptions

Before conducting the analysis, several basic principles or assumptions underlie regression analysis, several assumptions that underlie logistic regression analysis, namely [8].

1. Does not assume a linear relationship between the response variables and the predictor variables.
2. Predictor variables do not have to be normally distributed.
3. The response variable does not require the assumption of homogeneity for each level of the predictor variable or the variance does not have to be the same in each category.
4. The measurement scale on the response variable is discrete or binary (success/failure) and the predictor variable does not require an interval measurement scale.
5. Using probability sampling, which is a sampling technique to provide equal opportunities for each member of the population to be selected as a member of the sample.
6. Observation variables are measured without errors (valid and reliable measurement instruments) meaning that the variables studied can be observed directly.

4.3 Integrated cluster analysis method with logistic regression analysis

The linkage used in this study is the Average Linkage and the measurement of the distance between clusters using the Euclidean distance. Determination of the number has been determined in advance, namely as many as 2 and 3 groups. The Average Linkage method is based on the average distance. The table of the number of members in each Cluster in the Integrated Cluster Analysis method with regression analysis is presented in **Table 1**.

From **Table 1** it can be seen that there are 71 customers in Cluster 1 with 3 groups, 15 customers in Cluster 2, and 14 customers in Cluster 3. While many members with 2 groups in Cluster 1 as many as 93 customers and in Cluster 2 as many as 7 customers. The selection of the best linkage and model validity is by choosing the model that has the largest total R^2 , as shown in the equation, which can be briefly seen in **Table 2** as follows.

Based on **Table 2** the logistic regression analysis model with cluster integration with 3 groups has the greatest total determination value so that logistic regression

Cluster	Number of cluster members	
	3 Groups	2 Groups
1	71	93
2	15	7
3	14	—

Table 1.
Number of members of each cluster average linkage method on integrated cluster analysis method with logistics regression analysis.

	R^2 adjusted of Y_1	R^2 adjusted of Y_2	Total R^2 adjusted
3 Groups	0.4258	0.8492	0.8923
2 Groups	0.3852	0.8129	0.8667

Table 2.
Adjusted values R^2 for each integrated cluster analysis model with logistics regression analysis.

analysis with cluster integration with 3 groups is the best model compared to 2 groups. The total determination value of 89.23% is considered very good to describe the model.

Based on **Table 2** the adjusted R^2 value of the Cluster integration regression analysis with 3 groups resulted in an adjusted R^2 value of 0.4258 meaning that the variables of age, work experience, and loan to value were able to explain the diversity of credit collectibility variables of 42.58%, while 57 The other, 41% is influenced by variables outside the model. The value of R^2 adjusted Cluster inte-
gration logistic regression analysis with 3 groups resulted in an R^2 adjusted value of 0.8492, meaning that the variables of age, work experience, and loan to value were able to explain the diversity of credit collectibility variables of 84.92%, while 13.08. The other percentage is influenced by variables outside the model. The coefficient of total determination of the Cluster integration logistic regression analysis model with 3 groups is 0.8923, so it can be concluded that the diversity of data that can be explained by the model is 89.23% while the remaining 10.17% is explained by vari-
ables outside the model.

The results of R^2 the adjusted integrated cluster in logistic regression analysis with 3 groups having the highest adjusted R^2 value. If the average variables of each Cluster are compared, it is found that most of Cluster 2 has the highest average value compared to other Clusters, so Cluster 2 is high. While Cluster 1 has the lowest average value compared to other Clusters, so Cluster 1 is low. The average value for each cluster is presented in **Table 3**.

Based on **Table 3**, it can be seen that most of the customers are 39 years old in the low cluster, 37 years old in the high cluster, and 38 years old in the medium

Variable	Average		
	Cluster 1: low cluster	Cluster 2: high cluster	Cluster 3: medium cluster
Age (X_1)	39.507	37.333	38.571
Work experience (X_2)	39.930	193.867	107.571

Table 3.
Average value and each cluster in integrated cluster analysis model with logistic regression analysis.

cluster. The work experience of customers in the low cluster is mostly for 40 months, the high cluster is mostly for 194 months, while in the medium cluster mostly for 108 months.

Integrated Cluster Analysis method with Logistic Regression Analysis with 3 groups that separate each data set optimally. Then the model formed is like Eq. (2) as follows.

$$\pi(x) = \frac{\exp(-0,027x_1 - 0,041x_2 + 0,850y_1 - 0,374D_1x_1 - 0,006D_1x_2 + 9,971D_1y_1 + 0,090D_2x_1 + 0,026D_2x_2 - 1,559D_2y_1)}{1 + \exp(-0,027x_1 - 0,041x_2 + 0,850y_1 - 0,374D_1x_1 - 0,006D_1x_2 + 9,971D_1y_1 + 0,090D_2x_1 + 0,026D_2x_2 - 1,559D_2y_1)} \quad (2)$$

Low cluster ($D_1 = 0$ and $D_2 = 0$) can be seen in Eq. (3).

$$\pi(x) = \frac{\exp(-0,027x_1 - 0,041x_2 + 0,850y_1)}{1 + \exp(-0,027x_1 - 0,041x_2 + 0,850y_1)} \quad (3)$$

High cluster ($D_1 = 1$ and $D_2 = 0$) can be seen in Eq. (4).

$$\pi(x) = \frac{\exp(-0,401x_1 - 0,047x_2 + 10,821y_1)}{1 + \exp(-0,401x_1 - 0,047x_2 + 10,821y_1)} \quad (4)$$

Medium cluster ($D_1 = 0$ and $D_2 = 1$) can be seen in Eq. (5).

$$\pi(x) = \frac{\exp(0,063x_1 - 0,015x_2 - 0,709y_1)}{1 + \exp(0,063x_1 - 0,015x_2 - 0,709y_1)} \quad (5)$$

5. Integrated cluster with discriminant analysis

5.1 Discriminant analysis

Discriminant analysis is a multivariate analysis that functions to model the relationship between a categorical response variable and one or more quantitative predictor variables [9]. Discriminant analysis can be used as a grouping method because it produces a function that can distinguish between groups. The function is formed by maximizing the distance between groups. If the response variable or categorical data consists of only two groups, it is called a Two-Group Discriminant Analysis model, whereas if the group consists of more than two categories it is called Multiple Discriminant Analysis. Discriminant analysis has two assumptions that must be met, namely the assumption of multivariate normality, and the assumption of homogeneity of the variance matrix.

According to [6], discriminant analysis is included in the multivariate dependence method. The model can be written as in Eq. (6).

$$y_i = \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_p X_{pi} \quad (6)$$

where,

y_i : the response variable is categorical or nominal data on the i -th observation unit

X_{pi} : the p -explanatory variable on the i -th observation unit

β_p : the coefficient of the p -th discriminant function

i : 1, 2, 3, ..., n

5.2 Integration of cluster analysis with discriminant analysis of dummy variable approach

Integration of Cluster Analysis with Discriminant Analysis The Dummy Variable Approach in this study combines cluster analysis with discriminant analysis. Integrating cluster analysis with discriminant analysis can be done by using dummy variables obtained from cluster results. Many clusters formed are used as categories, then used as dummy variables.

An integrated cluster model with discriminant analysis can be written in Eq. (7).

$$y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_p x_{pi} + D_1 \beta_{p+1} x_{1i} + D_1 \beta_{p+2} x_{2i} + \dots + D_1 \beta_{p+q} x_{pi} + D_2 \beta_{p+q+1} x_{1i} + D_2 \beta_{p+q+2} x_{2i} + \dots + D_2 \beta_{p+2q} x_{pi} + \dots + D_q \beta_{p+q+1} x_{1i} + D_q \beta_{p+11+2} x_{2i} + \dots + D_q \beta_{p+qq} x_{pi} \quad (7)$$

where,

y_i : response variable at the i -th observation unit

x_{pi} : the p -explanatory variable on the i -th observation unit

β_p : the coefficient of the p -th discriminant function

D_q : q -th dummy variable

p : number of explanatory variables

q : the number of clusters formed is reduced by 1

i : 1, 2, 3, ..., n

If the research variables used are 3 and the number of clusters is 2, then the integrated cluster model with multiple discriminant analysis can be written as in Eq. (8).

Common models:

$$y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + D_1 \beta_4 x_{1i} + D_1 \beta_5 x_{2i} + D_1 \beta_6 x_{3i} \quad (8)$$

Cluster 1 ($D_1=0$)

$$y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} \quad (9)$$

Cluster 2 ($D_1=1$)

$$y_i = (\beta_1 + \beta_4) x_{1i} + (\beta_2 + \beta_5) x_{2i} + (\beta_3 + \beta_6) x_{3i} \quad (10)$$

5.3 Model efficiency

Efficiency can be seen based on three criteria, namely model accuracy, sensitivity, and specificity. Accuracy measures how correctly a diagnostic test identifies and excludes a certain condition, in other words, accuracy is used to measure the goodness of the model. In diagnostic tests, the terms sensitivity and specificity are also known. Sensitivity and specificity in diagnostic tests is a measure of the ability to correctly identify objects under reality [10]. The difference is that sensitivity measures the positive group while specificity measures the negative group. To get the value of accuracy, sensitivity, and specificity can use the Confusion Matrix as follows (Table 4).

$$\begin{aligned} \text{Accuracy} &= \frac{a + d}{a + b + c + d} \\ \text{Sensitivity} &= \frac{a}{a + c} \\ \text{Specificity} &= \frac{d}{b + d} \end{aligned} \quad (11)$$

Actual	Prediction	
	Z_1	Z_0
Z_1	a	b
Z_0	c	d

Table 4.
Confusion matrix.

5.4 Implementation of integrated cluster with discriminant analysis

For example, there are secondary data regarding homeownership loans obtained from Bank X in Indonesia, where the variables studied are age, credit period, loan to value, and credit collectibility status. The collectibility status of the credit used consists of two categories, namely the collectibility of current and non-current loans. The variables of age and credit period are in hours, while the loan to value is in proportion units. Therefore, it is necessary to standardize before conducting data analysis.

When using an integrated cluster with discriminant analysis, the first thing we have to do is perform a cluster analysis to get a dummy variable. In cluster analysis, it is not necessary to test assumptions because cluster analysis is included in exploratory analysis. If the results of the cluster analysis are n clusters, then the dummy variables formed are $n-1$ variables. The analysis used is hierarchical cluster analysis with the average linkage method with Euclidean distance. The determination of the number of clusters is determined based on the Silhouette value. Silhouette values for each of the many clusters can be seen in **Table 5**.

Based on **Table 5**, the largest Silhouette value is in many clusters 2. So that the optimum number of clusters is 2. The results of cluster analysis are obtained in cluster 1 consisting of 71 customers, and cluster 2 consisting of 29 customers. Thus, the dummy variable formed is 1 dummy variable. If the object (customer) is included in cluster 2, we assume that the object is 1 in the dummy variable. Meanwhile, if the object is included in cluster 1, we assume that the object is 0 in the dummy variable.

After obtaining the dummy variable, the next step is to test the assumptions in discriminant analysis. Testing for multivariate normality using the Shapiro-Wilk test on predictor variables, and testing the homogeneity of the covariance matrix using the Box M test. of 0.9917 (> 0.05). So it can be concluded that the data already meet the assumptions of multivariate normality and homogeneity of the variance matrix.

Next is to analyze the data using an integrated cluster with discriminant analysis. Based on the analysis carried out, an integrated cluster model was obtained with the following discriminant analysis:

Number of clusters	Silhouette value
2	0.4491
3	0.3915
4	0.2912
5	0.2811

Table 5.
Cluster analysis silhouette results.

$$y_i = 0,0838x_{1i} + 0.0606x_{12i} - 0,0241x_{3i} + 0,0569D_1x_{1i} + 0,0358D_1x_{2i} - 0,0752D_1x_{3i} \tag{12}$$

Cluster 1 ($D_1 = 0$)

$$y_i = 0,0838x_{1i} + 0.0606x_{12i} - 0,0241x_{3i} \tag{13}$$

Cluster 1 ($D_1 = 1$)

$$y_i = 0,1407x_{1i} + 0.0964x_{12i} - 0,0993x_{3i} \tag{14}$$

Based on the above equations, it can be interpreted that the coefficient of age and credit term is positive, meaning that the higher the age and credit term, the greater the possibility that customers in cluster 1 and cluster 2 have current credit collectability. On the other hand, loan-to-value has a negative coefficient, so if the value increases, it will increase the possibility of customers having non-current credit collectability. The variable that most influences credit collectability in cluster 1 and cluster 2 is age which has the largest discriminant coefficient value. The value of classification accuracy, sensitivity, and specificity in the integrated cluster analysis method with discriminant analysis can be seen in **Table 6** below.

Based on **Table 6**, the results of the classification accuracy are 84%, which means that the model correctly classifies as many as 84 customers out of 100 customers. Sensitivity of 84% means that customers belonging to the current category can be classified correctly by the model as many as 60 of 71 customers. The specificity of 16% means that customers belonging to the non-current category can be classified correctly by the model as many as 5 out of 29 customers.

6. Regression analysis with dummy variable

6.1 Regression analysis

The method that describes how big the relationship between variables is a regression analysis method. Regression analysis is divided into two, namely simple regression analysis and multiple regression analysis. Simple regression analysis is an analysis involving one predictor variable and one response variable, while multiple regression analysis is a regression analysis involving several predictor variables and one response variable. The regression analysis has several classical assumptions based on Gauss-Markov theory that must be met, namely the relationship between variables is correct, predictor variables are fixed or non-stochastic, homogeneity of variance, non-autocorrelation, error normality, non-multicollinearity [11].

6.2 Regression analysis with dummy variables

There are many ways to create a regression model with qualitative predictor variables, one of which is to use regression with dummy variables. The dummy variable is a variable used to obtain an estimator in a regression model involving

	Percentage
Classification accuracy	84%
Sensitivity	84%
Specificity	16%

Table 6.
Value of classification accuracy, sensitivity, and specificity.

qualitative predictor variables [12]. There is no difference in the assumptions underlying the regression with or without a dummy variable, this is because the addition of a dummy variable will be the same as the addition of a predictor variable in general.

There are several rules for coding dummy variables, for example by using binary code (0, 1). For example, there is a qualitative predictor variable with two categories (category 1 and category 2), then the qualitative variable can be defined in the dummy variable as shown in the following equation:

$$D = \begin{cases} 1, & \text{for category 1} \\ 0, & \text{for other} \end{cases} \quad (15)$$

The regression model with dummy variables can be expressed in the following equation:

$$Y_i = \beta_0 + \beta_1 D + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon_i \quad (16)$$

Information:

Y_i : the value of the response variable at the i -th observation.

X_i : the value of predictor i -th variable.

β : regression model parameter.

ε_i : Random error at i -th observation.

i : index for observation ($i = 1, 2, \dots, n$).

Dummy variables can be entered into the regression model in three different ways, namely:

1. Dummy variable as intercept component
2. Dummy variable as slope component
3. Dummy variables as components of intercept and slope

6.3 Application of regression analysis with dummy variables

From the available data, namely Y = willingness to pay, X_1 = dummy variable with category 1 being income in one family that is not combined, while category 2 is income in one family combined. X_2 is Service Quality, X_3 is Environment and X_4 is Fairness.

The regression model formed is $Y = b_0 + b_1 D + b_2 X_2 + b_3 X_3 + b_4 X_4$ (**Figure 1**)

Based on the regression analysis performed, the regression model with dummy variables is obtained as follows:

$$Y = 0.54088 + 0.08676 D + 0.1579X_2 + 0.4309X_3 + 0.2545X_4 \quad (17)$$

In this model, it is possible to know the difference in interest in paying creditors whose income is combined with income that is not combined.

6.4 Assumptions of regression analysis with dummy variable

6.4.1 Non multicollinearity

Multicollinearity is a problem in regression which means that the predictor variables correlate. A good regression model is a data that there is no

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.54088    0.23671   2.285 0.024540 *
dat$X11       0.08676    0.08096   1.072 0.286571
dat$X2        0.15794    0.07425   2.127 0.035992 *
dat$X3        0.43093    0.05248   8.211 1.08e-12 ***
dat$X4        0.25451    0.07078   3.596 0.000515 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3269 on 95 degrees of freedom
Multiple R-squared:  0.684,    Adjusted R-squared:  0.6707
F-statistic: 51.42 on 4 and 95 DF,  p-value: < 2.2e-16
```

Figure 1.
Output R.

multicollinearity problem. Multicollinearity checks can use the VIF value, where if the VIF value is <10 then there is no multicollinearity problem. In the data used, the VIF value for all variables is less than 10 so that the data is used to fulfill the assumption of non-multicollinearity.

6.4.2 Normality error

The assumption of normality of error is an assumption that requires that the error must be normally distributed with a mean of 0 and a variance σ^2 . Testing for normality of errors can use the Shapiro Wilk test.

H0: normal distribution error

H1: error is not normally distributed

$\alpha = 5\%$

Based on the normality test, a p-value of 0.91 was obtained, which means that the error was normally distributed. So that the assumption of normality error is met.

6.4.3 Non autocorrelation

The non-autocorrelation assumption test aims to find out whether some observations have correlated errors or not. If there is covariance and the correlation between errors is not equal to zero, then this can be said as a violation of assumptions. The non-autocorrelation assumption test method can be done using the Durbin Watson method. Based on the analysis conducted using the Durbin Watson test, a p-value of 0.6132 was obtained, which means that the data met the non-autocorrelation assumption.

6.4.4 Homoscedasticity

The assumption of homoscedasticity invariance indicates that as the average increases, the variance should remain constant, but there is a possibility that an increase in the average value causes the variance value to also increase, so it is necessary to test the assumption of homogeneity of variance. Assumption testing is done so that the estimator results obtained are efficient. Testing the assumption of homoscedasticity can use the Brusch Pagan method.

Based on the analysis conducted using the Brusch Pagan test, a p-value of 0.130 (less than 0.05) was obtained, which means that the data met the assumption of homoscedasticity.

6.5 Parameter significance test

a. Simultaneous test

$$H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$$

$$H_1: \text{there is at least one } \beta_i \neq 0$$

$$\alpha = 5\%$$

Based on the analysis obtained a p-value of 0.000 which means that there is at least one significant regression coefficient.

b. Partial test

$$H_0: \beta_i = 0$$

$$H_1: \beta_i \neq 0$$

$$\alpha = 5\%$$

Based on the analysis, it was found that three regression coefficients have a p-value of less than 0.05. The three regression coefficients are the coefficients of the variables X2 (Quality of Service), X3 (Environment), and X4 (Fairness). This means that Service Quality, Environment, and Fairness have a significant effect on Willingness to Pay.

6.6 Model interpretation

The model obtained and has fulfilled all the assumptions of regression analysis with dummy variables is as follows:

$$y = 0.54088 + 0.08676 D + 0.1579x_2 + 0.4309x_3 + 0.2545x_4 \quad (18)$$

In this model, it is possible to find out the difference in interest in paying creditors whose income is combined with income that is not combined. Based on the model obtained, the coefficient of the dummy variable is 0.08676, which means that when the incomes of creditors in one family are combined, the willingness to pay will be greater than those of creditors whose income is not combined. The estimated regression coefficient for the variable X2 (Quality of Service) is 0.1579, which means that the better the bank's service quality, the willingness to pay for credit also increases. Then for the estimation of the regression coefficient for the X3 (Environmental) variable, an estimate of 0.4309 is obtained, which means that the better the creditor's environmental conditions, the willingness to pay credit will also increase. The same thing also happened to the variable X4 (Fairness) where the estimated regression coefficient was 0.2545, which means that if the bank institution is fairer, creditors will also be more interested in paying.

7. Conclusion

The use of cluster analysis in statistical modeling will greatly facilitate the capture of the diversity of objects so that objects with the same characteristics can be grouped into the same group. This will be useful in classification methods such as discriminant analysis. Because in one group, objects will be more homogeneous, while between groups has a high diversity. So, the novelty in this chapter is the

perspective of the dummy variable technique where the number of categories in the dummy variable is determined by the number of clusters formed from the results of cluster analysis. This will then be continued on statistical modeling which is able to help researchers to divide objects into several groups according to the characteristics of each object by minimizing the diversity within the group.

Conflict of interest

The authors declare no conflict of interest.

Author details

Adji Achmad Rinaldo Fernandes*, Solimun and Nurjannah
Departement of Statistics, University of Brawijaya, Malang City, Indonesia

*Address all correspondence to: fernandes@staff.ub.ac.id

IntechOpen

© 2022 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/3.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

References

- [1] Stattrek.com. Dummy Variables in Regression. 2021. Available from: <https://stattrek.com/multiple-regression/dummy-variables.aspx>
- [2] Displayr.com. What are Dummy Variables?. 2019. Available from: <https://www.displayr.com/what-are-dummy-variables/>
- [3] Skrivanek S. The Use of Dummy Variables in Regression Analysis. Powell, OH: More Steam, LLC; 2009
- [4] Artaya IP. Analisa Regresi Linier Berganda Metode Dummy Banyak Kriteria. 2019. DOI: 10.13140/RG.2.2.13471.41122
- [5] Fernandes AAR. Metode Statistika Multivariat Pemodelan PERSamaan Struktural (SEM) Pendekatan WarPLS. Malang: UB Press; 2017
- [6] Johnson RA, Wichern DW. Applied Multivariate. Analysis. Upper Saddle River. NJ: Prentice-Hall; 2007
- [7] Gudono. Analisis Data Multivariat Edisi Pertama. Yogyakarta: BPFE; 2011
- [8] Tatham RL, Hair JF, Anderson RE, Black WC. Multivariate Data Analysis. New Jersey: Prentice Hall; 1998
- [9] Wong HB, Lim GH. Measures of Diagnostic Accuracy: Sensitivity, Specificity, PPV and NPV. Proceedings of Singapore Healthcare. 2011;20(4):316-318
- [10] Gujarati D. Ekonometri Dasar: Terjemahan Sumarno Zein. Erlangga: Jakarta; 2003
- [11] Nawari. Analisis Regresi dengan MS Excel 2007 SPSS 17. Jakarta: PT Elex Media Komputindo; 2010
- [12] Le Cessie S, Van Houwelingen JC. Logistic regression for correlated binary data. Journal of the Royal Statistical Society: Series C (Applied Statistics). 1994;43(1):95-108