Modeling Governance KB with CATPCA to Overcome Multicollinearity in the Logistic Regression

I. Khikmah\(^1\), H Wijayanto, and U D Syafitri

Department of Statistics, Faculty of Mathematics and Natural Science, Bogor Agriculture University, Bogor, Indonesia.

Corresponding author: aisyah.salsabila17@gmail.com

Abstract. The problem often encountered in logistic regression modeling are multicollinearity problems. Data that have multicollinearity between explanatory variables with the result in the estimation of parameters to be bias. Besides, the multicollinearity will result in error in the classification. In general, to overcome multicollinearity in regression used stepwise regression. They are also another method to overcome multicollinearity which involves all variable for prediction. That is Principal Component Analysis (PCA). However, classical PCA in only for numeric data. Its data are categorical, one method to solve the problems is Categorical Principal Component Analysis (CATPCA). Data were used in this research were a part of data Demographic and Population Survey Indonesia (IDHS) 2012. This research focuses on the characteristic of women of using the contraceptive methods. Classification results evaluated using Area Under Curve (AUC) values. The higher the AUC value, the better. Based on AUC values, the classification of the contraceptive method using stepwise method (58.66%) is better than the logistic regression model (57.39%) and CATPCA (57.39%). Evaluation of the results of logistic regression using sensitivity, shows the opposite where CATPCA method (99.79%) is better than logistic regression method (92.43%) and stepwise (92.05%). Therefore in this study focuses on major class classification (using a contraceptive method), then the selected model is CATPCA because it can raise the level of the major class model accuracy.

1. Introduction

In the field of modeling studies often found an association between variables, whether the relationship between the response variable with the explanatory variables, the relationship between the response variable, as well as the relationship between the explanatory variables. The relationship between the response variable will cause the relationship between remnant. This will result in inter-residual graphs cannot describe the actual state of the data, so that interpretation to be incorrect. However, the relationship between the response variable can still be controlled when the number of degrees of freedom and a lot of large remnants [2].

Relationships between more than two explanatory variables are often referred to as multicollinearity. In the case of a linear relationship between the explanatory variables in the regression parameter estimates lead to be inappropriate, although the coefficient determination generated great and significant. Multicollinearity between explanatory variables in the logistic regression model parameter estimation will cause to be invalid. This is because the variance covariance matrix that will be formed to be interconnected, so that parameter estimation is not appropriate [1].

Some researchers did multicollinearity detection before further analysis. If the explanatory variables are numeric variables, then to detect multicollinearity using correlation matrices, variance influence...
factor (VIF), and the eigenvalues of the correlation matrix. However, if the explanatory variables are categorical variables, then to detect multicollinearity using association analysis.

Multicollinearity techniques for handling classified into two methods, the method of removal (discard the explanatory variables are correlated) and variable selection method. In general, the variable selection method used is the best subset, the method, and the backward stepwise regression method. But the stepwise method and VIP still not up to handle their multicollinearity, so use some method development. In the case of linear data, handling multicollinearity can be resolved by methods PLS-VIP, methods PLS-BETA, LASSO and stepwise regression [3]. Bastien [4] examined using PLS-LR and Aguilena[5] using PCLR. Both of these methods are good enough to handle multicollinearity in the logistic regression. However, if the explanatory variables in the form of nominal and ordinal data (nonlinear), Kemalbay[1] developed a method Categorical Principal Component Analysis (CATPCA). This study will compare the methods of stepwise regression and CATPCA method by using case studies of governance KB modeling with logistic regression.

2. Material and Methods

2.1. Categorial Principal Component Analysis (CATPCA)

The goal of traditional principal component analysis (PCA) is to reduce the number of \( m \) variables to a smaller number of \( p \) uncorrelated variables called principal components which account for the variance in the data as much as possible. Since PCA is suitable for continuous variables which are scaled at the numerical level of measurement such that interval or ratio and it also assumes linear relationship among variables, it is not an appropriate method of dimension reduction for categorical variables. Alternatively, categorical (also known as nonlinear) principal components analysis (CATPCA) has been developed for the data given mixed measurement level such that nominal, ordinal or numeric which may not have a linear relationship with each other. For categorical variables, CATPCA uses optimal scaling process which transforms the category labels into numerical values while the variance accounted for among the quantified variables is maximized [6]. We refer to [7] for a historical review of CATPCA using optimal scaling. For continuous numeric variables, the optimal scaling process is as the traditional case. Suppose we have measurement of \( n \) individuals on \( m \) variables given with an \( n \times m \) observed scores matrix \( H \) where each variable is denoted by \( jX \), \( j=1, \ldots, m \) that is the \( j \)th column of \( H \). If the variable \( jX \) are of nominal or ordinal measurement level, then a nonlinear transformation called optimal scaling is required where each observed scores transformed into category quantification given by:

\[
q_j = \varphi_j(X_j)
\]

where \( Q \) is the matrix of category quantifications. Let \( X \) be the \( n \times p \) matrix of object scores, which are the scores of the individuals on the principal components, obtained by CATPCA. The object scores are multiplied by a set of optimal weights which are called component loadings. Let \( A \) be an \( m \times p \) matrix of the component loadings where the \( j \)th column is denoted by \( jA \). Then the loss function for minimization of difference between original data and principal components can be given as follows:

\[
L(Q, A, S) = n^{-1} \sum_{j=1}^{m} \text{tr}(q_j a_j^T - S)(q_j a_j^T - S)
\]

2.2. Logistic Regression Multinomial

Logistic regression analysis is an analysis that is used to model the response variable \( Y \) that is based on the categorical explanatory variables \( X \) that are numerical and categorical. Logistic regression analysis requires several things (assumptions) about the nature of the data, namely:

1. The dependent variables should be categorical.
2. There is no significant correlation between the independent variables.
3. The number of observations for each variable should be adequate, and the overall sample size is large enough. The number of samples of at least 400 units to be able to get a good fitness model [8].

The conditional probability can be written:
$$E(y|x) = \frac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p)}{1 + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p)}$$
(3)

or

$$\pi(x) = \frac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p)}{1 + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p)}$$
(4)

Equivalently, logit has a linear equation as follows:

$$\text{Logit } \left[ \frac{\Pi(x)}{1-\Pi(x)} \right] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$$
(5)

which is a bridge function (link function) of the explanatory variables [9].

Testing parameters simultaneously performed with a likelihood ratio test (likelihood ratio test). A likelihood ratio test (likelihood ratio test) is obtained by comparing the functions of all variables explanatory Log-likelihood with Log-likelihood function without the explanatory variables. With the hypothesis that will be tested as follows:

$$H_0: \beta_j = \beta_{j-1} = \ldots = \beta_p = 0 \text{ (no explanatory variables that influence the response variable)}$$

$$H_1: \text{At least there's the one with } \beta_j \neq 0, \ j = 1, 2, \ldots, p \text{ (at least one explanatory variables that influence the response variable).}$$

The test statistic G formulated as follows

$$G = -2 \ln \frac{l_0}{l_1}$$
(6)

where

- $l_0$: likelihood without explanatory variable
- $l_1$: likelihood with explanatory variable

G statistic follows a chi-square distribution with $df = (c-1)(r-1)$. For accept or reject $H_0$, then compare the G value and $X^2$. Reject $H_0$ if $G > X^2_{(p-c), \alpha}$ or reject $H_0$ if $p-value < \alpha$ [8]. Partial testing is done by using the Wald test. The hypothesis tested is:

$$H_0: \beta_j = 0 \text{ (} X_j \text{ explanatory variables did not significantly influence the response variable)}$$

$$H_1: \beta_j \neq 0 \text{ (} X_j \text{ explanatory variables did significantly influence the response variable),}$$

for $j = 1, 2, \ldots, p$.

Formula for the Wald test statistic is

$$W = \frac{\hat{\beta}_j}{SE(\hat{\beta})}$$
(7)

3. Methods

3.1. Data

Data were used in this research is secondary data. Data were a part of data Indonesian Demographic Health Survey (IDHS) in 2012 obtained from Demographic and Health Surveys (DHS). The unit of observation in the form of household, namely the productive age group of women aged 15-49 years as many as 29 882 people. Variables used a categorical variable by 8 variables. Variables used are shown in Table 1.

| Table 1. Variables | Category     |
|---------------------|--------------|
| Contraceptive methods | 1 = not used, 2 = used |
| Age in 5-year Groups (x) | 1 = <30 years old |
3.2. Analysis Method
Steps of data analysis performed in this study are as follows: (1) Explore the data to get a general overview of user IDHS data 2012. (2) Modeling tool selection without contraception with logistic regression using stepwise regression method and CATPCA. In this step will be performed with logistic regression modeling while maintaining the association between the explanatory variables. So it will be visible difference modeling results with and without maintaining the association between the explanatory variables. The model will be established by equation 6 and 7. (3) Handling associations between explanatory variables with CATPCA. (4) Logistic regression modeling tool selection method contraception stepwise regression. (5) Logistic modeling tool selection method contraception CATPCA. (6) Comparing CATPCA with stepwise regression and classification table.

4. Results and Discussion

4.1. Classification Table
After testing the suitability of the models that have been obtained, then it was performed the accuracy measurement of the classification multinomial logistic regression model that has been formed. Classification accuracy results can be seen in Table 2.

| Performance classification | Result |
|---------------------------|--------|
| Accuracy                  | 61.54  |
| Sensitivity               | 92.43  |
| Specificity               | 14.50  |
| AUC                       | 57.39  |

Based on Table 2, it can be seen that the accuracy of classification models using logistic regression models good enough for 61.54%. AUC values of 57.39% indicate that the model's accuracy is good enough. The percentage accuracy of the classification model quite well. A value large enough sensitivity that is equal to 92.43%, which means that the accuracy of classification for methods of contraception would be appropriate but still errors may occur a minority who will be classified to the majority class. Classification of a logistic regression model with a stepwise method is done by entering the variables most associated with Y is the age category, the wife of education, knowledge of contraceptives, the number of living children, living and working place today, so we get Table 3.
Based on Table 3, it can be seen that the accuracy of classification models using a stepwise logistic regression model with good enough for 60.80%. AUC values of 58.66% indicate that the model's accuracy is good enough. The percentage accuracy of the classification model quite well. A value large enough sensitivity that is equal to 92.05%, which means that the accuracy of classification for methods of contraception would be appropriate but still errors may occur a minority who will be classified to the majority class. With CATPCA, we get Table 4.

| Performance classification | Result |
|----------------------------|--------|
| Accuracy                   | 60.80  |
| Sensitivity                | 92.05  |
| Specificity                | 13.20  |
| AUC                        | 58.66  |

Table 4. Classification Performance with CATPCA Logistic Regression Model

| Performance classification | Result |
|----------------------------|--------|
| Accuracy                   | 61.16  |
| Sensitivity                | 99.79  |
| Specificity                | 2.34   |
| AUC                        | 57.39  |

Based on Table 4, it can be seen that the accuracy of classification models using a stepwise logistic regression model with good enough for 61.16%. AUC values of 57.39% indicate that the model's accuracy is good enough. The percentage accuracy of the classification model quite well. A value large enough sensitivity that is equal to 99.79%, which means that the accuracy of classification for methods of contraception would be appropriate but still errors may occur a minority who will be classified to the majority class.

Comparison Model

The model that has been obtained is then compared to the level of accuracy with AUC values. AUC values on stepwise logistic regression model (57.89) are higher than the logistic regression model (57.39) and CATPCA (57.89). This shows that the stepwise logistic regression model is more accurate than the logistic regression models and models CATPCA. In addition to comparing the AUC values obtained, may be considered sensitivity and specificity of each model. AUC value was composed by sensitivity and specificity. In this study, more sensitivity focused on value, because in this case, the use of contraceptive methods is more likely than not using a contraceptive method. According to Table 4, the model can improve sensitivity CATPCA big enough though specificity decreased. Sensitivity models CATPCA (99.79) is higher than the logistic regression model (92.43) and stepwise models (92.05). Thus, based on the sensitivity of the model to predict the data CATPCA minor class with a lot better.

5. Conclusions and Recommendations

This research concluded that Based on AUC values, the classification of the contraceptive method using stepwise logistic regression model is better than the logistic regression mode and CATPCA. While based on the value of sensitivity, the model CATPCA better than the logistic regression model and stepwise. In this study focuses on major class (the use of contraceptive methods), CATPCA can raise specificity major class model. In general, in the condition data is not balanced method CATPCA less well in the accuracy of the classification, but were able to classify correctly. This research recommends that In the actual case, there are four categories in which the two categories have a small percentage so that data is not balanced. So to further research needs to be treated assessment of the data is not balanced so as to optimize the performance of classification.

References

[1] Kemalbay Gand Ozlem BK2014 Procedia - Social and Behavioral Sciences 109 730 DOI: 10.1016/j.sbspro.2013.12.537
[2] Draper NR and Smith H 1998 *Applied Regression Analysis* 3rd Ed ISBN: 978-0-471-17082-2 (New York: John Wiley & Sons Inc)
[3] Chong I and Chi-Hyuck J 2005 *Chemometr Intell Lab* 78 103 doi:10.1016/j.chemolab.2004.12.011
[4] Bastien P, Vinzi V E and Tenenhaus M 2005 *Comput Stat Data An* 48 17 doi:10.1016/j.csda.2004.02.005
[5] Aguilena AM, Manuel E and Mariano JV 2006 *Comput Stat Data An* 50 1905 doi:10.1016/j.csda.2005.03.011
[6] Linting M and Van der Kooij A 2012 *J Pers Assess* 94 12. DOI:10.1080/00223891.2011.627965
[7] Gifi A 1990 *Nonlinear multivariate analysis* ISBN: 978-0-471-92620-7 (Chichester: John Wiley and Sons)
[8] Hosmer DW and Lemeshow S 2000 *Applied Logistic Regression* 2nd Ed. (New York: Wiley) DOI: 10.1002/0471722146
[9] Agresti A 2002 *Categorical Data Analysis* 2nd Ed. (New York: John Wiley & Sons)