Available online at www.sciencedirect.com

ScienceDirect

Procedia - Social and Behavioral Sciences 109 (2014) 730 – 736

Procedia
Social and Behavioral Sciences

2nd World Conference On Business, Economics And Management - WCBEM 2013

Categorical Principal Component Logistic Regression: A Case Study for Housing Loan Approval

Gülder Kemalbay [a] *, Özlem Berak Korkmazoğlu [a]

[a]*Yıldız Technical University, Davutpasa Campus, Faculty of Art&Science, Department of Statistics, 34220, Esenler, İstabul, Turkey*

Abstract

The logistic regression describes the relationship between a binary (dichotomous) response variable and explanatory variables. If there is multi collinearity among the explanatory variables, the estimation of model parameters may lead to invalid statistical inference. In this study, we have survey data for 2331 randomly selected customers which consists of highly correlated binary explanatory variables to model whether a customer's housing loan application has been approved or not. For this purpose, we present a categorical principal component analysis to deal with the multi collinearity problem among categorical explanatory variables while predicting binary response variable with logistic regression.

Keywords: Categorical principal component analysis, multicollinearity, binary data, logistic regression;

1. Introduction

Binary response variables frequently arise in many research areas especially for applications in the biomedical and social sciences. Logistic regression is a specific type of generalized linear models (GLM), a class of nonlinear regression models, which is commonly used when the response variable is binary and the explanatory variables are either continuous or discrete. When there is a multi collinearity problem among explanatory variables, the estimation of the logistic regression coefficients may lead to invalid statistical inference. Aguilera *et.al.* (2006) Proposed the principal component analysis by using a reduced set of principal components of the continuous explanatory variables as covariates of the logistic regression. Then, they checked the performance of the proposed model on a simulation study. For further discussion about using principal component analysis in logistic regression model, one can see Marx and Smith (1990), Camminatiello and Lucadamo (2010), among others.
The purpose of the present paper is to improve the accuracy of the estimation of logistic regression coefficients when we have a binary response variable and a large number of highly correlated categorical explanatory variables. For this purpose, we first employ a categorical principal component analysis to deal with multi collinearity problem among binary explanatory variables, and then use uncorrelated principal components instead of original correlated

* Corresponding Author: Gülder Kemalbay. Tel.: +90-0212-383-4429.
E-mail address: kemalbay@yildiz.edu.tr

doi:10.1016/j.sbspro.2013.12.537

variables to regress the binary response variable with logistic regression model. For numerical illustration of the proposed categorical principal component logistic regression model, we analyse a survey data to investigate the factors affecting the housing loan approval of a private bank in Turkey.

2. Association for Variables Measured at the Nominal Level

Measure of association quantifies the strength of a relationship between variables and also it helps to analyze an evidence of the cause and effect relationship. In bivariate case, the independent and dependent variables are taken as the cause and effect, respectively. The measure of association for variables measured at nominal level is a statistics based on the value of Chi-square. A few of the most widely used ones are: *Lambda* (λ), is a directional (asymmetrical) measure of association which provides us the strength of relationship between independent and dependent variables given by the formula $\lambda = (E_1 - E_2)/E_1$, where E_1 is the prediction error made when the independent variable is ignored and E_2 is the prediction error made when the independent variable is taken into account. This proportion explains the extent to which predictions of the dependent variable are improved by considering the independent variable. Since it is asymmetric, the value of the statistic depends on which variable is taken as independent. *Phi* (ϕ) is a symmetric measure of association for strength of relationship, which is appropriate for nominal-by-nominal data given in 2x2 contingency table. Its formula is expressed as $\phi = \sqrt{\chi^2/N}$, where χ^2 is Chi-square statistics. $\phi \in [0,1]$ As a greater than 0.30 indicates a strong relationship (Healey, 2012).

3. Categorical Principal Component Analysis

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 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 (Linting and Van der Kooij, 2012). We refer to Gifi (1990) for an 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 $\mathbf{H}$ where each variable is denoted by $\mathbf{X}_j$, j=1,...,m that is the j^{th} column of $\mathbf{H}$. If the variables $\mathbf{X}_j$ 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:

$$\mathbf{q}_j = \varphi_j(\mathbf{X}_j), \tag{1}$$

where $\mathbf{Q}$ is the matrix of category quantifications. Let $\mathbf{S}$ 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 $\mathbf{A}$ be $m \times p$ matrix of the component loadings where the j^{th} column is denoted by $\mathbf{a}_j$. Then the loss function for minimization of difference between original data and principal components can be given as follows:

$$L(\mathbf{Q},\mathbf{A},\mathbf{S}) = n^{-1}\sum_{j=1}^{m} tr(\mathbf{q}_j\mathbf{a}_j^T - \mathbf{S})^T(\mathbf{q}_j\mathbf{a}_j^T - \mathbf{S}), \tag{2}$$

where *tr* is the trace function, i.e. for any matrix $\mathbf{A}$, $tr(\mathbf{A}^T\mathbf{A}) = \sum_i \sum_j a_{ij}^2$. Consequently, the CATPCA is performed by minimizing the least-squares loss function given in the equation (2) in which the matrix $\mathbf{X}$ is replaced by the

matrix **Q**. The loss function is exposed to some restrictions. First, $\mathbf{q}_j^T\mathbf{q}_j = n$, that is transformed variables are standardized to solve the indeterminancy between $\mathbf{q}_j$ and $\mathbf{a}_j$. This standardization indicates that $\mathbf{q}_j$ contains *z*-scores and yields that the component loadings in $\mathbf{a}_j$ are correlations among transformed variables and principal components. The object scores are restricted by $\mathbf{S}^T\mathbf{S} = n\mathbf{I}$, where **I** is the identity matrix, to avoid the trivial solution. However, the object scores are centered, i.e. $1^T\mathbf{S} = 0$, where 1 is a vector of ones. These restrictions imply that the columns of **S** are orthonormal *z*-scores (Linting *et. al.,* 2007). The minimization of restricted loss function given in (2) is obtained by means of an Alternating Least Squares (ALS) algorithm (Gifi, 1990).

4. Logistic Regression with Binary Response

Let Y be a binary response variable, which is coded as 0 or 1, referred to as absence or presence, respectively. Then the logistic regression model is given as follows:

$$\pi(x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}, \tag{3}$$

$\pi(x)$Represents the conditional mean of Y given x, i.e. $E(Y|x)$. The value of response variable given x can be expressed as $y = \pi(x) + \varepsilon$, ε is the error term. If $y = 1$, then $\varepsilon = 1 - \pi(x)$ with probability $\pi(x)$ and if $y = 0$, $\varepsilon = -\pi(x)$ with probability $1 - \pi(x)$. Therefore, ε follows a binomial distribution with mean 0 and variance $\pi(x)[1-\pi(x)]$. A transformation of $\pi(x)$ which is called *logit function* is required:

$$g(x) = \ln\left[\frac{\pi(x)}{1-\pi(x)}\right] = \beta_0 + \beta_1 x. \tag{4}$$

The unknown parameters are estimated by the method of maximum likelihood estimation with given likelihood function for $\boldsymbol{\beta} = (\beta_0, \beta_1)$ given as $L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \pi(x_i)^{y_i}[1-\pi(x_i)]^{1-y_i}$.

4.1. Fitting Logistic Model with Binary Explanatory Variables

Let us consider the interpretation of the coefficients for logistic regression model with the case where explanatory variables are at the nominal level of measurement. Assume that X is coded either 0 or 1. Then the difference between logit function when x=1 and x=0 is given as $g(1) - g(0) = \beta_1$. To interpret this result, a measure of association called odds ratio (OR) is required:

$$OR = \frac{\pi(1)/[1-\pi(1)]}{\pi(0)/[1-\pi(0)]} = e^{\beta_1}. \tag{5}$$

Odds ratio provides an approximation how much more likely or unlikely it is for the response variable to occur among those with x = 1 than among those with x = 0. For details, one can see Hosmer and Lemeshow (2000).

5. Numerical Results

As a case study, a survey data for 2331 randomly selected customers which is collected by a private bank is analysed whether a customer's housing loan application has been approved or not by using categorical principal component logistic regression. The dependent variable, housing loan approval, is a binary variable coded 1 for approved applicants and 0 for disapproved applicants. The effects of some explanatory variables on housing loan approval are investigated. The explanatory variables used are: Age(X_1; 1 for age<35, 2 for age>35), Education (X_2; 1 for graduate, 2 for others), Marital Status (X_3; 1 for single, 2 for married), Gender (X_4; 1 for male, 2 for female), Turkish Lira Time Deposit Account Ownership (X_5; 1 for yes, 2 for no), Turkish Lira Deposit Account Ownership

(X_6; 1 for yes, 2 for no), Foreign Currency Time Deposit Account Ownership (X_7; 1 for yes, 2 for no), Foreign Currency Deposit Account Ownership (X_8; 1 for yes, 2 for no), Gold Account Ownership (X_9; 1 for yes, 2 for no), Repo (X_{10}; 1 for yes, 2 for no), Bond (X_{11}; 1 for yes, 2 for no), Mutual Fund Ownership (X_{12}; 1 for yes, 2 for no), Stock (X_{13}; 1 for yes, 2 for no), Treasury Bill (X_{14}; 1 for yes, 2 for no) and Life Insurance (X_{15}; 1 for yes, 2 for no). Since all of the explanatory variables are at the nominal level of measurement, Lambda and Phi coefficients are used for measure of association. Among others, highly associated pairs of variables are presented in the Table 1.

Table 1. The Measure of Association Coefficients for Some Explanatory Variables

Variables	**Directional Measure**		**Symmetric Measure**	
Nominal by Nominal	Lambda	Approx. Sig.	Phi	Approx. Sig.
Dependent*İndependent (a)	λ	p	ϕ	p
$X_5 * X_7$	0.982	0.000	0.987	0.000
$X_6 * X_8$	0.663	0.000	0.809	0.000
$X_{10} * X_{12}$	0.339	0.000	0.566	0.000
$X_{11} * X_{14}$	0.861	0.000	0.927	0.000
$X_{13} * X_{15}$	0.360	0.000	0.576	0.000

a Dependent and Independent variables are considered for calculation of Lambda.

One possible way to check of multicollinearity for nominal measured variables would be use to Lambda and Phi coeffients. As seen in Table 1, these coefficients indicate a strong relationship between some explanatory variables. When multicollinearity occurs among variables, the estimated logistic regression coefficients may be inaccurate, in other words it reduces the predictive power of the model. Therefore, the categorical principal component analysis is performed to reduce the observed variables to a number of uncorrelated principal components. The CATPCA model is summarized in Table 2:

Table 2. Model Summary of CATPCA

Dimension	**Variance Accounted For**		
	Cronbach's Alpha	Total (Eigenvalue)	% of Variance
1	0.691	2.813	18.754
2	0.543	2.028	13.517
3	0.482	1.818	12.118
4	0.306	1.401	9.338
5	0.246	1.298	8.651
Total	0.957(a)	9.357	62.377

a Total Cronbach's Alpha is based on the total Eigenvalue.

In CATPCA, the eigenvalues are obtained from the correlation matrix between the quantified variables and the variance account for of the first p components is maximized simultaneously over nonlinear transformed variables. The eigenvalues are complete summary measure which provide the variance accounted for by each principal component. If the original correlated variables form two or more sets, then more than one principal component is required to summarize the variables. The eigenvalues shows that how accomplished this summary is.

In this analysis, the reason why we ignore the dimensions higher than five is that their contribution is very little to the total variance accounted for. Also, another reason is to prefer principal components correspondig eigenvalues are greater than 1. But this criterion mat not be always optimal. The-five dimensional CATPCA on the housing loan approval data ensures the largest eigenvalue of 2.813, providing that 18.754% of the variance in the transformed variables is explained by the first component. The eigenvalue of the second component is 2.028, providing that its percentage of variance accounted for is 13.517%, and other components account for as much as possible of the remaining variance, respectively. Thus, all of the components account for a substantial percentage 62.377% of the total variance in the transformed variables.

In Table 3, the component loading matrix is given. Component loadings are equal to Pearson correlation coefficient among principal components and quantified variables. The principal components in CATPCA are weighted sums of the quantified variables. The object scores corresponding to each individuals on the components are obtained.

Table 3. The Matrix of Component Loadings

Variable	Dimension				
	1	2	3	4	5
X_1	-0.038	-0.075	0.160	-0.281	0.562
X_2	0.188	0.135	-0.093	0.270	-0.174
X_3	-0.032	0.024	0.003	-0.189	0.762
X_4	-0.097	0.067	-0.244	-0.045	0.487
X_5	0.422	0.826	-0.038	-0.329	-0.060
X_6	0.278	-0.136	0.875	-0.109	-0.032
X_7	0.420	0.830	-0.037	-0.326	-0.061
X_8	0.301	-0.090	0.870	-0.123	-0.019
X_9	0.357	0.252	-0.053	0.250	0.108
X_{10}	0.611	-0.300	-0.156	-0.112	-0.052
X_{11}	0.698	-0.452	-0.268	-0.254	-0.009
X_{12}	0.576	-0.091	-0.124	-0.009	-0.045
X_{13}	0.496	0.157	0.047	0.615	0.220
X_{14}	0.677	-0.451	-0.262	-0.245	-0.010
X_{15}	0.469	0.076	0.113	0.626	0.243

Therefore, we reduce the dimension of the logistic regression model to avoid multicollinearity by using five number of principal components as explanatory variables. In Table 4, we present some goodness of fit measures for logistic regression model.

Table 4. Goodness of Fit Statistics for Logistic Regression

Model Summary				Hosmer and Lemeshow Test		
Step	-2 Log Likelihood	Cox &Snell R Square	Nagelkerke R Square	Chi-square	df	Sig.
1	839.632	0.484	0.756	9.175	8.000	0.328

One of these measures is Hosmer-Lemeshow test shows that the model ensures better fit than a null model with no explanatory variables. If the test statistic is not significant, then it means that the model adequately fits the data. According to Table 4, Hosmer-Lemeshow goodness of fit test statistics is not significant (p-value is 0.328) which implies that the estimated model fit the data at a convenient level.

To summarize the logistic regression model, there are some approximations for coefficient of determination R^2, called pseudo R^2. However these are not goodness-of-fit tests but rather measure srength of association. One of them is Cox & Snell R^2, indicates 48.4% of the variation in the response variable is explained by the model. However, there is more reliable measure, Nagelkerke R^2 indicates a strong relationship of 75.6% between explanatory variables and the prediction. Both pseudo measures tends to be lower than the traditional R^2. In addition to goodness of fit statistics, the classification results presented in Table 5 tell us how many of the cases where the observed values of the response variable have been correctly predicted.

Table 5. Classification Table [(a), (b)]

Observed			Predicted		
			Housing Loan Approval		Percentage
			No	Yes	Correct
Step 1	Housing Loan	No	613	112	84.6
	Approval	Yes	95	1511	94.1
Overall Percentage					91.1

a Constant is included in the model, b The Cut Value is 0.500.

The results presented in Table 5 show that how accurate the model is at predicting whether a customer's application for housing loan has been approved or not. Then, 84.6% of applicants to housing loan and 94.1% of the non-applicants are correctly classified. The overall percantage is 91.1% considerably high percentage of customers are classified for housing loan approval.
In the following table, we present the estimation results for logistic regression in the case where principal components are used as explanatory variables.

Table 6. Estimation Results for Logistic Regression

	B	S.E.	Wald	df	Sig.	Exp(B)	95,0% C.I.for EXP(B)	
	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper
COMP1	-3.495	0.191	336.335	1.000	0.000	0.030	0.021	0.044
COMP2	-2.849	0.195	213.992	1.000	0.000	0.058	0.040	0.085
COMP3	-0.193	0.076	6.500	1.000	0.011	0.824	0.711	0.956
COMP4	-0.828	0.095	75.559	1.000	0.000	0.437	0.362	0.527
COMP5	1.395	0.090	239.288	1.000	0.000	4.037	3.383	4.818
Constant	2.496	0.158	249.709	1.000	0.000	12.128		

The **B** column is the parameter estimation of the logistic regression model. The Wald statistic is a common way to test the significance of estimation for each explanatory variables. If this statistics is significant, then we reject the null hypothesis in logistic regression as the variable contributes significantly to the estimation. It is seen that all principal components has significant contribution to the estimation. Then, the logistic regression model for this study can be expressed as follows:

$$\frac{e^{(2.496-3.495Comp1-2.849Comp2-0.193Comp3-0.828Comp4+1.395Comp5)}}{1+e^{(2.496-3.495Comp1-2.849Comp2-0.193Comp3-0.828Comp4+1.395Comp5)}} \quad . \quad (6)$$

However, **Exp (B)** column provides odds ratio, which provides the relative importance of the explanatory variables on the response variable's odds. For example, **EXP(B)** value corresponding to the Comp5 is 4.037 which means that Comp5 is approximately 4 times as important as in determining the decision for approval of the housing loan application.

6. Conclusions

CATPCA can be used as an alternative to the widely known linear methods of dimension reduction for the data given mixed measurement level such that nominal, ordinal or numeric which may not have linear relationship with each other. As clear from the classification table, using principal components as the explanatory variables provides rather high correct classification rate of housing loan approval, indicating that an appropriate strategy to model this type of variables is selected. 84.6% of applicants to housing loan, 94.1% of the non-applicants are correctly

classified using our logistics regression model with 91.1% overall correct classification rate. Consequently, the presented categorical principal component logistic regression is a convenient method to improve the accuracy of logistic regression estimation under multicollinearity among categorigal explanatory variables while predicting binary response variable.

References

Aguilera, M.A., Escabias, M., & Valderrama, J.M. (2006). Using principal components for estimating logistic regression with high-dimensional multicollinear data. *Computational Statistics & Data Analysis*, 50, 1905-1924.

Camminatiello, I., & Lucadamo, A. (2010). Estimating multinomial logit model with multicollinear data. *Asian Journal of Mathematics and Statistics*, 3(2), 93-101.

Gifi, A. (1990). Nonlinear multivariate analysis. *John Wiley and Sons*. Chichester, England.

Healey, J. (2012). The Essentials of Statistics: A Tool for Social Research. *Wadsworth Publishing*, 3th edition, USA.

Hosmer, W.D., & Lemeshow, S. (2000). Applied Logistic Regression. *Wiley-Interscience Publication*. 2nd edition, New York.

Hosmer, D.W., Hosmer, T., Le Cessie, S., & Lemeshow, S. (1997). A comparison of goodness-of-fit tests for the logistic regression model. *Statistics in Medicine*, 16, 965–980.

Linting, M., Meulman, J. J., Groenen, P. J. F., & Van der Kooij, J. J. (2007). Nonlinear principal components analysis: introduction and application. *Psychological Methods*, 12, 336-358.

Linting, M., & Van der Kooij, A. (2012). Nonlinear principal components analysis with CATPCA: a tutorial. *Journal of Personality Assessments*, 94(1), 12-25.

Marx, B.D., & Smith, E.P. (1990). Principal component estimators for generalized linear regression. *Biometrika*, 77(1), 23–31.