Nonlinear Principal Component Logistic Regression: Impact of Morning Shows on Fashion Products Consumer

Dr Junaid Sageer Siddiqu, Dr Bushra Shamshad, Zara Omar

Department of Statistics, University of Karachi

Abstract
Logistic regression describe the relationship between binary response variable and explanatory variables. In this study, online survey is conducted to find the impact of morning shows on the viewers who purchase the fashion products and clothing brands because of celebrity endorsement. The sample of 1275 responses were collected to analyze the influence on viewers of morning shows who purchase the fashion products. The study using logistic regression model is apply to find the impact of morning shows on the consumer of fashion product.

Keywords:
Logistic regression, nonlinear principal components analysis, viewers impact survey data, scale construction.

1. Introduction

PCA is a useful technique in multivariate analysis and it can be extended for categorical data as well. Binary response are common in the field of medical and social sciences and for this purpose NLPCA is effective method. Generalized linear model for categorical data can be customize under logistic regression which is nonlinear regression model. The model proposed by Aguilera et.al.2006, principal component analysis by using a reduced set of principal components of the continuous explanatory variables as covariates of the logistic regression. The model is used to estimate the logistic regression coefficients. To avoid the multi-collinearity among the random variable, NLPCA is efficient method for it. In this research paper, parameters are estimated by NLPCA logistic regression model coefficients, as the data consist a binary response variable and highly correlated categorical explanatory variables. For numerical illustration of the NLPCA logistic regression model, the analysis of a survey data of 1275 observations has been investigate to check the impact of morning shows on the fashion product consumer in Pakistan.

2. Literature Review:

The history of NLPCA is the extension of the term of multivariate analysis called *multiple correspondence analysis*, a literal translation of Benz´ecri'*s L'analyse des correspondances* (*multiples*) (Benz´ecri, 1973, 1992). This history can be traced in the work of Fisher (1948), Guttman (1941), Burt (1950), and Hayashi (1952), among others, and in the rediscoveries since the 1970s (among others, see Benz´ecri, 1992; de Leeuw, 1973 Greenacre, 1984; Lebart, Morineau, & Warwick, 1984; Saporta, 1975; Tenenhaus & Young, 1985). The class of techniques is also known under the names *dual scaling* (Nishisato, 1980, 1994) and *homogeneity analysis* (Gifi, 1981/1990). In the course of its development, the technique has been given many different interpretations. In the original formulation of Guttman (1941), the technique was described as a principal components analysis of qualitative (nominal) variables. There is also an interpretation as a form of generalized canonical correlation analysis (Lebart & Tabard, 1973; Masson, 1974; Saporta, 1975), based on earlier work by Horst (1961a, 1961b), Carroll (1968), and Kettenring (1971).

Another major motivation to optimal scaling was given by work in the area of nonmetric multidimensional scaling (MDS), pioneered by Shepard (1962a, 1962b), Kruskal (1964), and Guttman (1968). In MDS, a set of proximities between objects is approximated by a set of distances in a low-dimensional space, usually Euclidean. Optimal scaling of the proximities was originally performed by monotonic regression; later on, spline transformations were incorporated (Ramsay, 1982). Since the so-called nonmetric breakthrough in MDS in the early 1960s, optimal scaling has subsequently been integrated in multivariate analysis techniques that hitherto were only suited for the analysis of numerical data. Some early contributions include Kruskal (1965), Shepard (1966), and Roskam (1968).

In the 1970s and 1980s, psychometric contributions to the area became numerous. Selected highlights from the extensive psychometric literature on the subject include de Leeuw (1973); Kruskal and Shepar (1974); Young, de Leeuw, and Takane (1976); Young, Takane, and de Leeuw (1978); Nishisato (1980); Heiser (1981); Young (1981); Winsberg and Ramsay
(1983); Van der Burg and de Leeuw (1983); Van der Burg, de Leeuw, and Verdegaal (1988); and Ramsay (1988).

Attempts at systematization resulted in the ALSOS system by Young et al. (1976), Young et al. (1978), and Young

Manuscript received March 5, 2019
Manuscript revised March 20, 2019

(1981) and the system developed by the Leiden "Albert Gifi" group. Albert Gifi's (1990) book, *Nonlinear Multivariate Analysis*, provides a comprehensive system, combining optimal scaling with multivariate analysis, including statistical developments such as the bootstrap. Since the mid-1980s, the principles of optimal scaling have gradually appeared in the mainstream statistical literature (Breiman & Friedman, 1985; Buja, 1990; Gilula & Haberman, 1988; Hastie et al., 1994; Ramsay, 1988). The Gifi system is discussed among traditional statistical techniques in Krzanowski and Marriott (1994).

3. Nonlinear Principal Component Analysis

PCA is used to reduce the number of variables to the smaller number of variables with no multi-collinearity among them, which account for the variance in the data as much as possible. PCA is a useful term for continuous variables which has interval or ratio scaling of measurement. The basic assumption for PCA is to have linear relationship between the variables which is not possible in every case. The NLPCA is free from the assumption of linearity and which is a useful for categorical data which has nominal scales of measurement. For categorical variables, NLPCA 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). The basic reference of NLPCA is Gifi (1990), defines an historical review of NLPCA using optimal scaling.

The p variables on k individuals given with an $p \times k$ observed scores matrix H where each variable is denoted by X_l , $l = 1, \ldots, k$ that is the l^{th} column$H$. If the variables X_l are of nominal or ordinal measurement level, the transformation of nonlinear variables called optimal scaling. It is required for each observed scores to transform into category quantification given by:

$$q_l = \gamma_l X_l \tag{1}$$

Categorical quantification is define using matrix Q. The matrix S of order $p \times n$ of object scores, which are the scores of the individuals on the principal components, obtained by NLPCA. The object scores are multiplied by a set of optimal weights which are called component loadings. Let A be $k \times n$ matrix of the component loadings where the l^{th} column is denoted by a_l . Then the loss function for minimization of difference between original data and principal components can be given as follows:

$$L(Q, A, S) = p^{-1}(\sum_{l=1}^{k} tr(q_l {a_l}^T - S)^T (q_l {a_l}^T - S)) \tag{2}$$

where tr is the trace for any matrix A, $tr(A^T A) = \sum_{i,j} {a_{ij}}^2$. The NLPCA is performed by minimizing the least-squares loss function given in the above equation in which the matrix X is replaced by the Q.

4. Logistic Regression with Binary Response

Let Y be a binary response variable, which is coded as 0 or 1, referred to as yes or no, respectively. Then the logistic regression model is given as follows:

$$\pi(y) = \frac{e^{a+by}}{1-e^{a+by}} \tag{3}$$

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

$$g(y) = \log \frac{\pi(y)}{1-\pi(y)} = a = by \tag{4}$$

The unknown parameters are estimated by the method of maximum likelihood estimation with given likelihood function for $\beta = (a, b)$ given as $L(\beta) = \prod_{i=1}^{n} \pi(y_i)^{x_i}[1 - \pi(y_i)]^{x_i}$ .Consider the interpretation of the coefficients for logistic regression model with the case where explanatory variables are at the nominal level of measurement. To interpret the results obtain from logistic regression, a measure of association called odds ratio (OR) is required. 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 $x = 0$.

5. Data Analysis

Data is collected by online survey, the sample of 1275 observations are used to find the impact of morning show and celebrity endorsement on the sale of branded cloths in the market of Pakistan. The variables which is considered in this research is Gender (Male=A, Female=B), Fashion in your view (A=Looking cool, B=Looking sober and classy, C=Going with the current trends and D=A way to express my inner self), Expense on fashion products and Clothing Brands (A=Gul Ahmed, B=AlKaram Studio,C=J.,D= Sana Safinaz, E=Nishat Linen) . The research is based on viewers of morning shows who are influence in purchasing fashion products and clothing brands by the celebrity

endorsement in the shows. Results from sample indicates that the highly correlated the variables are and shows the impact of fashion brands and clothing is influences by the morning shows and celebrity endorsements.

Table 1: The value and p-value of Wald test for measuring the association among the variables

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 0	Constant	.360	.057	39.981	1	.000	1.433

The above table shows that the intercept-only model is ln(odds) = 0.360 and the expression has predicted odds [Exp(B)] = 1.433, which is the deciding value. The Wald chi square test also shows that the constant term is not zero which means the minimum value for the model is 0.360. It is showing that the significance of variable dependency is high and significant.

Table 2: The model summary with -2log likelihood, Cox & Snell R Square and Nagelkerke R Square

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	66.876a	.46	.62

Model shows that the variation is well explained in the dependent variable. The above table shows that the -2 Log Likelihood is 66.87, which defines that the model is good fit and has higher impact of the variables. The model summary contains both Cox & Snell R Square and Nagelkerke R Square values, which are methods to obtained the explained variation and can be interpreted like R^2 in a multiple regression. The values are also known as *pseudo* R^2 values and has lower value than multiple regression. It can be interpreted in the same manner as R^2. The model has explained 46 % to 62% of variation in the dependent variable, respectively. Nagelkerke R^2 is a modification of Cox & Snell *R2*.

Table 3: The model parameters with degree of freedom and 95 % Confidence Interval for Exp(B)

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
								Lower	Upper
Step 1a	Gender(1)	.084	.128	.428	1	.513	1.087	.846	1.397
	Fashion in you view	-.369	.074	25.010	1	.000	.691	.598	.799
	Expense	.430	.000	1.017	1	.013	1.000	1.000	1.000
	Clothing brand	.164	.046	12.847	1	.000	1.179	1.077	1.289
	Constant	.643	.240	7.213	1	.007	1.903		

a. Variable(s) entered on step 1: Gender, Fashion in you view, Expense, Clothing brand.

The Wald test is used to determine statistical significance for each of the independent variables using the value from significance value. The results shows that *Fashion in your view* (p = .000) , *Expense* (p = .013) and *Clothing Brand* (p=0.000) are significant in the model, but *Gender*(p = .513), did not add significantly to the model. The information given by the table is used to predict the probability to predict of an event occurring based on a one unit change in an independent variable when all other independent variables are kept constant. Confidence interval of β is the range of values that the values have 5% of error in the calculation or shows the confident that each odds ratio lies within .05 error part. The setting of 95% means that there is only a p < .05 that the value for the obtain parameter is not correct. The variables *Clothing Brands and Fashion In Your View* shows a high impact on the purchase of fashion products. The logistic regression equation will be as:

$$Log\ Y = 0.643 + 0.084Gender - .369Fashion\ in\ your\ view + .430Expense + .164Clothing\ Brand$$

The above equation shows that the least value of the log Y is 0.643. Gender, Expense and Clothing Brands has positive impact on model with $0.084Gender$ value, if Gender will increase by 1 unit log y will increase by 0.084 unit, as $0.430Expense$, if Expense will increase by 1 unit log y will increase by 0.43 unit and lastly 0.164Clothing Brand if Clothing Brands will increase by 1 unit log y will increase by 0.164 unit. The variable Fashion in your view has negative impact on the model with, $-0.369Fashion\ in\ your\ view$, if fashion in your view will increase by 1 unit model will decrease by 0.369 unit.

Table 4: Goodness of fit test using Hosmer and Lemeshow method

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	12.422	8	.310

The goodness of fit test using Hosmer Lemeshow test is also showing that the model is significant and adequately fitting the data. The null hypothesis is that the model is a 'good enough' fit to the data (as $p>.05$). Model shows that the data is appropriately fitted and has significant p-value. The overall facts and figure shows that the morning shows has strong influence on the consumer of fashion products who are also shows viewers and they made decisions under the influence of celebrity endorsement and brands impact due to these shows.

6. Conclusions

The model used by the Aguilera (2006) has been used for the estimation of model parameters and shows the significant efficiency of the model as well as avoiding the multi-collinearity among the variables. Logistics regression can be an effective techniques using NLPCA method of estimation of parameters rather than simple logistic regression analysis which can also influence by the multi-collinearity and causes lacking of significance of statistics inferential estimation. It is very clear from the tabulated values that using NLPCA logistic regression as the model, the explanatory variables provides rather highly correlated variables using an appropriate strategy to model this type of variables is selected. Products purchase by the consumer is influenced by the brand endorsement of morning shows is modeled using nlpca logistics regression model with 62.1% overall correct classification rate. Consequently, the presented nonlinear principal component logistic regression is a convenient method to improve the accuracy of logistic regression estimation under multi-collinearity among categorical explanatory variables while predicting binary response variable.

References

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

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

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

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

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

[6] 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.

[7] Korkmazoglu, G. K. (2014). Categorical Principal Component Logistic Regession: A case study for Housing Loan Approval. 2nd World Conference On Business, Economics And Management, (pp. 730-736).

[8] 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.

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

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