MTISD 2008 – Methods, Models and Information Technologies for Decision Support Systems
Università del Salento, Lecce, 18-20 September 2008

Estimating multinomial logit model with multicollinear data

Ida Camminatiello
Department of "Matematica e Statistica"
University of Neaples Federico II
camminat@unina.it

Antonio Lucadamo
Department of "Scienze Economiche e Metodi Quantitativi"
University of "Piemonte orientale"
alucadam@unina.it

Abstract: *The multinomial model is used to study the dependence relationship between a categorical response variable with more than two categories and a set of explicative variables. In presence of multicollinearity, the estimation of the multinomial model parameters becomes inaccurate. To solve this problem we develop an extension of Principal Component Logistic Regression (PCLR), model proposed by Aguilera et al. (2006). Finally a case study illustrates the advantages of the method.*

Keywords: Multinomial regression, multicollinearity, principal component regression.

1. Introduction

The logistic regression model is used to predict a binary response variable in terms of a set of explicative ones. As many authors have stated (Ryan, 1997), the logistic model becomes unstable when strong correlation among predictors (multicollinearity) exists. To solve this problem Marx (1992) introduces iteratively reweighted partial least squares algorithm, Bastien *et al* (2005) propose partial least squares logit regression, Aguilera *et al* (2006) present Principal Component Logistic Regression (PCLR), Vágó and Kemény (2006) develop the ridge logistic regression.

When the response is measured on nominal scale with more than two categories, the multinomial regression is applied. In this paper we show, in presence of multicollinearity, the estimation of the multinomial model parameters becomes inaccurate because of the need to invert nearsingular and ill-conditioned information matrices. To provide an accurate estimation of the model parameters, we propose an extension of PCLR model. Finally, the performance of the proposed model is analyzed by developing a bootstrap procedure.

2. A solution to multicollinearity: Principal Component Multinomial Regression .

In order to improve the estimation of the multinomial model parameters (Ben-Akiva and Lerman, 1985) in the case of multicollinearity, we propose to use as covariates of the multinomial model a reduced number of Principal Components (PC) of the predictor variables.

Let be $\mathbf{X} = \lfloor \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p \rfloor$ a set of *p* quantitative independent variables, $\mathbf{y}$ a categorical response variable with more than two categories. The aim of Principal Components Analysis (PCA) is to find a set of uncorrelated latent variables $\mathbf{Z} = \lfloor \mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_p \rfloor$ which are linear combinations of the original variables $\mathbf{Z} = \mathbf{XV}$. The weight matrix $\mathbf{V} = \lfloor \mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_p \rfloor$ is built by the eigenvectors of the correlation matrix $\mathbf{R}$, whose elements are the correlation coefficients among the regressors. The basic properties of the analysis are:

1. the PC's are orthogonal;
2. the weights used to determine the PC's maximize the variance among the **x** variables, so the first $a < p$ PC's lead to a good approximated reconstruction of original matrix $\mathbf{X} = \mathbf{ZV}$

At this point, we can present the new approach: Principal Component Multinomial Regression (PCMR). At first step, PCMR creates the PC's of the regressors as described above. At second step the multinomial model is carried out on the subset of p PC's . The probability, for the individual "n", to choose the alternative "i" can be expressed in terms of all pc's as

$$\pi_n(i) = \frac{\exp\left\{\sum_{j=1}^{p}\sum_{k=1}^{p} z_{ik} v_{jk} \beta_j\right\}}{\exp\left\{\sum_{i=1}^{s}\sum_{j=1}^{p}\sum_{k=1}^{p} z_{ik} v_{jk} \beta_j\right\}} = \frac{\exp\left\{\sum_{k=1}^{p} z_{ik} \gamma_k\right\}}{\exp\left\{\sum_{i=1}^{s}\sum_{k=1}^{p} z_{ik} \gamma_k\right\}} \qquad (1)$$

where $z_{ik}, (i = 1, \ldots, n; k = 1, \ldots, p)$ are the elements of the PC's matrix, $\gamma_k = \sum_{j=1}^{p} v_{jk} \beta_j, (k = 1, \ldots, p)$ is the vector of coefficients to be estimated, β_j is the parameter vector expressed in function of original variables and s is the number of alternatives of the data set.
At third step, the number of PC's $a < p$, to be retained in the model, is chosen. The next paragraph discusses about the different tools for selecting the number of PC's. At fourth step, the multinomial model is carried out on the subset of $a < p$ PC's. Finally, the multinomial model parameters can be expressed in function of original variables in **X** matrix.

$$\mathbf{Z}_{(\mathbf{a})}\boldsymbol{\gamma}_{(\mathbf{a})} = \mathbf{X}\mathbf{V}_{(\mathbf{a})}\boldsymbol{\gamma}_{(\mathbf{a})} = \mathbf{X}\boldsymbol{\beta}_{(\mathbf{a})} \qquad \text{where } \boldsymbol{\beta}_{(\mathbf{a})} = \mathbf{V}_{(\mathbf{a})}\boldsymbol{\gamma}_{(\mathbf{a})} \qquad (2)$$

This leads to lower variance estimates of model parameters comparing to classical multinomial model. We estimate the variance of the estimated parameters of the multinomial model by bootstrap resampling. The bootstrap estimate of variance of the j estimated parameter, $S(\hat{\beta}^{J}_{(a)})$, is the empirical estimate calculated from m bootstrap values:

$$S(\hat{\beta}^{J}_{(a)}) = \frac{1}{m}\sum_{i}^{m}\left(\hat{\beta}^{J}_{i(a)} - \bar{\hat{\beta}}^{J}_{(a)}\right)^2 \quad \text{where} \qquad \bar{\hat{\beta}}^{J}_{(a)} = \frac{1}{m}\sum_{i}^{m}\hat{\beta}^{J}_{i(a)} \qquad (3)$$

2.1 Model calibration and validation

The number of PC's a is bounded from above by p, the number of **x** variables. Hence, the number of components should be chosen in the range $1 \le a \le p$. The number of PC's, a, to be retained in the model can be selected according to different tools.
The first possibility is to retain all the components, but the most used criteria are:

1. To Consider the PC's in their natural order and to stop when explained variability is about 75%
2. To Consider the PC's that correspond to eigenvalues bigger than one.

However, the dependence relationship between the response and the predictor variables is not taken into account. For this reason we decide to consider in our model all the PC's that influence the response variable.

To determine the goodness of the different criteria, we develop a bootstrap procedure and we use the bootstrap samples to estimate the parameters, both for the original matrix and for the PC's matrix. The bootstrap procedure shows, the best results are obtained, when the two following criteria are used at the same time:

1. to delete the components that are relatively unimportant as predictors of the original independent variables (**X**), i.e., the components having the eigenvalue lower than one should be dropped;
2. to delete the components that are not significant as predictors of the dependent variable (**y**).

3. A case study

Our study concerns the choice of the residence in the Zurich area. The data matrix is built after an household survey and it contains many variables about sociodemographic features of the households, the characteristics of their dwelling and housing price information (B✧rgle, 2006). We select, for the estimation of residential location choice, 17 variables (i.e. rent, distance from working place, distance from city center, etc), according to many available publications about similar problems and we estimate a Multinomial Logit Model (Bierlaire, 2003; Bierlaire, 2008).

The dataset shows independent variables strongly correlated. Several cells of correlation matrix are close to 1 and the condition index tends to infinity. As consequence, only six independent variables influence in significant manner the choice of the residence. However, we believe the other independent variables are important to characterize the several alternatives in unambiguous manner. Moreover the results obtained by bootstrap procedure on the original matrix are discouraging: the significant independent variables differ from a sample to another; the estimated parameter signs change from a sample to another.

In order to solve these problems we decide to apply the PCMR. As stated before, at first, we carry out the Principal Component Analysis, obtaining, for the observations, the coordinates on the factorial axes. Then we use as covariates of the multinomial model a reduced number of PC's of the predictor variables.

Different criteria are considered to decide the number of components to retain in the model. For an overview, we can look at the results in table 1, in which we compute the bootstrap estimate of variance of the estimated parameters using the formula (3). We consider here only the variables that are significant on the original dataset, but the results are similar for the other parameters.

Variables	Basic value	All pc's	PCMR(7)	PCMR(6)	PCMR(2)
Distance from working place	0.795633	0.034862	0.00767	0.01947	0.00264
Rent in CHF	0.101965	0.08315	0.00298	0.00262	0.000545
Density of open space (500 m radius)	0.000354	0.016107	0.048756	0.0026	0.00191
Density of young people (1 km radius)	4.390690	0.002317	0.068501	0.01596	0.00296
Sun index	0.119155	0.026059	0.034599	0.00974	0.00593
Driving time to city center	2.917186	0.456245	0.96759	0.00444	0.00417

Table 1. Variance of estimated parameters for the different methods.

In the first column there is the name of the variable, then, from second to sixth column we have the variance of the estimated parameters, calculated for the five different datasets: the original matrix; all PC's; the seven PC's that influence in significant manner the choice of the residence; the six PC's with eigenvalue bigger than one; the two components that are significant to explain the dependent variable and that have eigenvalue higher than one. It is easy to observe:

1. the variance of the estimated parameters computed on the original variables, is always bigger than the estimated variance calculated using the PCMR except than for the variable "open space".
2. lower variance estimates of model parameters (last column) are obtained retaining in the model only the significant components with eigenvalue bigger than one.

4. Conclusion and perspectives

In this paper we carried out a model for choice of the residence in the Zurich area. The data set showed independent variables strongly correlated, so the estimation of the multinomial model parameters presented a high variance. To solve the problem, we proposed to use as covariates of the multinomial model a reduced number of PC's of the predictor variables. In order to select the optimum PCMR model we considered and compared different methods for including PC's in the model. The method that considered only the significant components with eigenvalue bigger than one yielded lower variance estimates of model parameters.

The application showed that the proposed approach is a valid alternative on real data. However, an extensive simulation study is needed in order to verify the results and the method for selecting the PC's.

Finally, an extension of the other models, proposed in literature, for solving the problem of high-dimensional multicollinear data in the logit model could be interesting.

Bibliography

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

Bastien P., Esposito V., Tenenhaus M. (2005), PLS generalised linear regression, Computational Statistics & Data Analysis, 48 (1): 17-46.

Ben-Akiva M. E., Lerman S. R. (1985), Discrete Choice Analysis: Theory and Application to Travel Demand, MIT Press, Cambridge, Ma.

Bierlaire M. (2003), BIOGEME: A free package for the estimation of discrete choice models, Proceedings of the 3rd Swiss Transportation Research Conference, Ascona, Switzerland.

Bierlaire M. (2008), An introduction to BIOGEME Version 1.6, biogeme.epfl.ch.

B✧rgle M., (2006), Residential location choice model for the Greater Zurich area, Proceedings of the 6th Swiss Transport Research Conference, Ascona, Switzerland.

Marx B.D. (1992), A continuum of principal component generalized linear regression, Computational Statistics & Data Analysis, 13: 385-393.

Ryan T.P. (1997), Modern Regression Methods, Wiley, New York .

Vágó E., Kemény S. (2006), Logistic ridge regression for clinical data analysis (a case study), Applied ecology and environmental research 4 (2): 171-179.