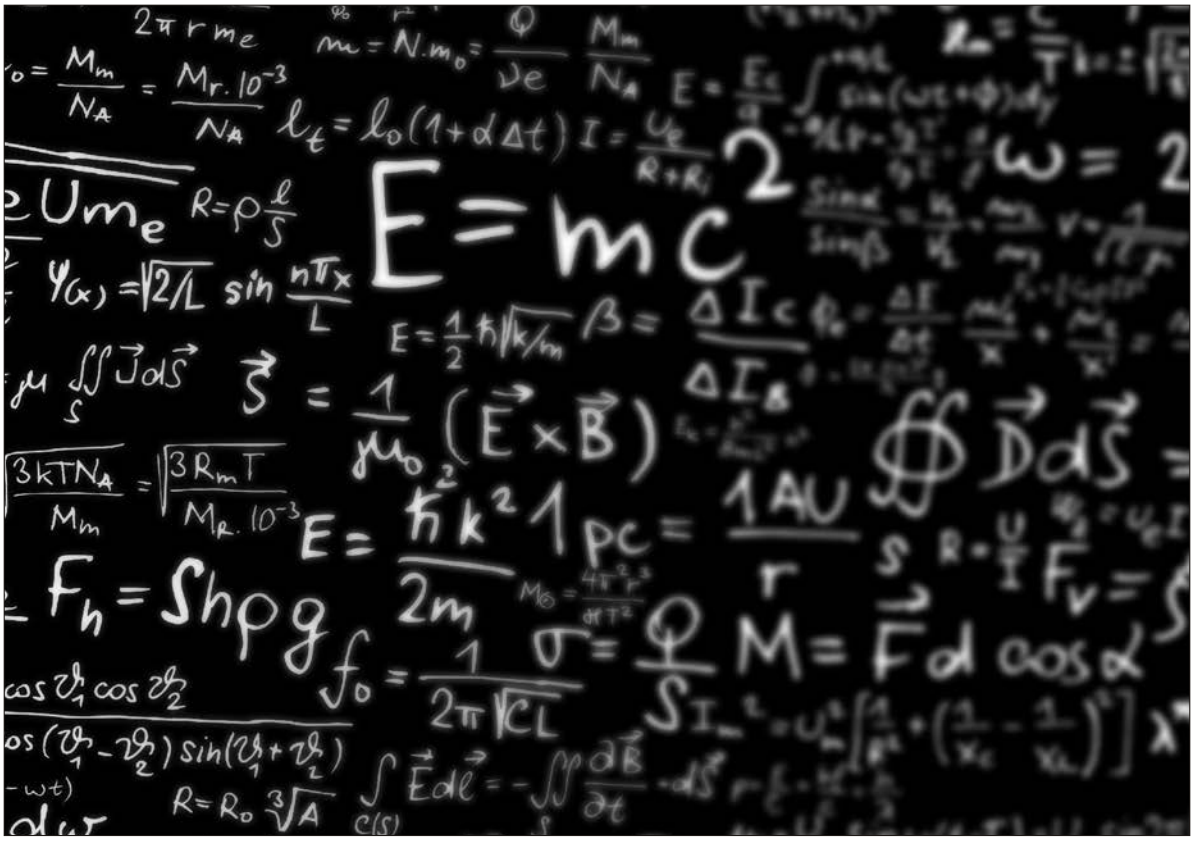

Asian Journal of Mathematics & Statistics

ISSN 1994-5418

science alert

ANSInet
an open access publisher
http://ansinet.com

Asian Journal of Mathematics and Statistics 3 (2): 93-101, 2010
ISSN 1994-5418

Estimating Multinomial Logit Model with Multicollinear Data

[1]I. Camminatiello and [1]A. Lucadamo
[1]Department of Matematica e Statistica,
University of Neaples Federico II, via Cinthia, 80126 Napoli, Italy
[2]Department of Scienze Economiche e Metodi Quantitativi,
University of Piemonte Orientale, Novara, Italy

Abstract: The multinomial logit 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 logit model parameters becomes inaccurate. To solve this problem we develop an extension of principal component logistic regression. Finally a simulation study illustrates the advantages of the method.

Key words: Multinomial logit regression, multicollinearity, principal component regression

INTRODUCTION

The regression analysis studies the statistical dependence of one or more dependent variables, Y, on one or more explanatory variables, X. All procedures used and conclusions drawn in a regression analysis depend on assumptions of a regression model. The most used model is the classic linear regression model and the most used method for estimating classic model parameters is the method of Ordinary Least Squares (OLS).

Under the classic assumptions this method has some attractive statistical properties that have made it one of most powerful and popular methods of regression analysis. However, OLS is not appropriate when strong correlation among predictors (multicollinearity) exists, so alternative methods to OLS have been developed.

The presence of multicollinearity may indicate that some explanatory variables are linear combinations of the other ones. Consequently, they do not improve explanatory power of a model and could be dropped from the model. In some situations, it is not feasible to use variable selection to reduce the number of explanatory variables or it is not desirable to do so.

An alternative way of dealing with unpleasant consequences of multicollinearity lies in biased estimation: we can sacrifice a small bias for a significant reduction in variance of an estimator. This observation motivates a whole class of biased estimators called shrinkage estimators.

There are various shrinkage methods that perform well under multicollinearity and that can possibly act as variable selection tools as well: the Ridge Regression (RR) (Hoerl and Kennard, 1970) and its modifications, Partial Least Squares (PLS) regression (Wold, 1966, 1985) Principal Components Regression (PCR) (Massy, 1965).

Corresponding Author: Ida Camminatiello, Department of Matematica e Statistica,
University of Neaples Federico II, via Cinthia, 80126 Napoli, Italia

Frank *et al.* (1993) compared the methods above mentioned. Although the results are conditional on the simulation design used in the study, the researchers indicate that RR, PCR and PLS have similar proprieties, give similar performance and are highly preferable to variable selection.

When the response is measured on nominal scale with more than two categories, the multinomial logit model (Luce, 1959; McFadden, 1974) is applied.

In this study, 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 a new strategy based on PCR. Then, the performance of the proposed model is analyzed by developing a simulation study. In order to validate the results obtained from each simulation we apply bootstrap procedure.

A SOLUTION TO MULTICOLLINEARITY: PRINCIPAL COMPONENT MULTINOMIAL REGRESSION

The binary logit model is used to predict a binary response variable in terms of a set of explicative ones. When the response is measured on nominal scale with more than two categories, the multinomial logit model is applied. Both the models become unstable when there is multicollinearity among predictors (Ryan, 1997).

To improve the estimation of the binary logit model parameters, Marx (1992) introduced iteratively reweighted partial least squares algorithm, Bastien *et al.* (2005) proposed partial least squares logit regression, Aguilera *et al.* (2006) presented Principal Component Logistic Regression (PCLR), Vágó and Kemény (2006) developed the ridge logistic regression.

Following the approach proposed by Aguilera *et al.* (2006), we propose to use as covariates of the multinomial logit model a set of orthogonal variables, linear combination of original ones, in order to provide an accurate estimation of the parameters.

Here, we describe Principal Component Analysis (PCA) and multinomial logit regression, then we illustrate our idea.

Principal Component Analysis

Principal Component Analysis (PCA) is a multivariate technique that transforms a number of correlated variables into a (smaller) number of uncorrelated variables with maximum variance, called Principal Components (PC). The first principal component accounts for as much of the variability in the data as possible and each succeeding component accounts for as much of the remaining variability as possible.

Let $X = \lfloor x_1,x_2,...,x_p \rfloor$ be a set of p quantitative independent variables, y a categorical response variable with more than two categories. The aim of PCA is to find a set of uncorrelated latent variables $Z = \lfloor z_1,z_2,...,z_p \rfloor$ which are linear combinations of the original variables $Z = XV$. The weight matrix $V = \lfloor v_1,v_2,...,v_p \rfloor$ is built by the eigenvectors of the covariance or correlation matrix. These matrices can be calculated from the data matrix X. The covariance matrix contains scaled sums of squares and cross products. The correlation matrix is similar to the covariance matrix but first the variables, i.e., the columns, have been standardized. For the reasons which are beyond the scope of this paper, it is often preferable to perform the analysis on correlation matrix R, whose elements are the correlation coefficients among the independent variables. The basic proprieties of the analysis are:

- The PC's are orthogonal
- 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 $X = ZV^T$

Multinomial Logit Regression

Multinomial logit regression is the simplest model in discrete choice analysis when more than two alternatives are in a choice set. It is derived from utility-maximizing theory that states that consumer chooses the alternative which maximizes his utility. Obviously not all the attributes of the alternatives will be observed and for this reason the utility is divided in two parts:

- D_{ib} is the systematic part of the utility that the individual i receives by a generic alternative b
- ε_{ib} is the random part and summarizes the contribution of unobserved variables (Ben-Akiva and Lerman, 1985). The probability to select a specific alternative c for the individual i is then:

$$\pi_i(c) = \Pr(D_{ic} + \varepsilon_{ic} \geq D_{ib} + \varepsilon_{ib}) \quad \forall \quad c \neq b; b = 1,\ldots,s \tag{1}$$

where, D_{ic} is the systematic part of the utility that the individual i receives by the alternative c, ε_{ic} is the disturbance and *s* is the number of alternatives.

If we assume that the disturbances are independent and identically extreme value distributed (Marschak, 1960) we obtain the multinomial logit model. The probability can be then expressed as follows:

$$\pi_i(c) = \frac{\exp(\mu D_{ic})}{\sum_{b=1}^{s} \exp(\mu D_{ib})} \tag{2}$$

The term μ is a scale parameter and it can be normalized to 1; furthermore, if the systematic part of the utility is linear in the parameters, we have (Train, 2003):

$$\pi_i(c) = \frac{\exp(x_{ij}\beta_{jc})}{\sum_{b=1}^{s} \exp(x_{ij}\beta_{jb})} \tag{3}$$

where, x_{ij}, (i = 1,...,n,; j = 1,...,p) are the elements of the X matrix and β_{jb} are the parameters to be estimated.

Principal Component Multinomial Regression

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 set of p PC's. The probability, for the individual i, to choose the alternative c can be expressed in terms of all PC's as:

$$\pi_i(c) = \frac{\exp\left\{\sum_{j=1}^{p}\sum_{k=1}^{p} z_{ik} v_{kj} \beta_{jc}\right\}}{\left\{\sum_{b=1}^{s}\exp\sum_{j=1}^{p}\sum_{k=1}^{p} z_{ik} v_{kj} \beta_{jb}\right\}} = \frac{\exp\left\{\sum_{k=1}^{p} z_{ik} \gamma_{kc}\right\}}{\left\{\sum_{b=1}^{s}\exp\sum_{k=1}^{p} z_{ik} \gamma_{kb}\right\}} \quad (4)$$

Where:

z_{ik}, (i = 1,...,n; k = 1,...,p) = The elements of the PC matrix
v_{kj}, (j = 1,...,p) = The elements of the transposed matrix V^T
$\gamma_{kb} = \sum_{j=1}^{p} v_{kj}\beta_{jb}, (b = 1,\ldots, s)$ = The coefficients to be estimated
β_{jb} = The parameters expressed in function of original variables and
S = 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. The probability, for the individual i, to choose the alternative c can be expressed in terms of a PC's as:

$$\pi_i^{(a)}(c) = \frac{\exp\left\{\sum_{j=1}^{p}\sum_{k=1}^{a} z_{ik} v_{kj} \beta_{jc}^{(a)}\right\}}{\left\{\sum_{b=1}^{s}\exp\sum_{j=1}^{p}\sum_{k=1}^{a} z_{ik} v_{kj} \beta_{jb}^{(a)}\right\}} = \frac{\exp\left\{\sum_{k=1}^{a} z_{ik} \gamma_{kc}^{(a)}\right\}}{\left\{\sum_{b=1}^{s}\exp\sum_{k=1}^{a} z_{ik} \gamma_{kb}^{(a)}\right\}} \quad (5)$$

Where:

$\gamma_{kb}^{(a)} = \sum_{j=1}^{p} v_{kj}\beta_{jb}^{(a)}$ = The coefficients to be estimated on the subset of a PC's
$\beta_{jb}^{(a)}$ = The PCMR parameters obtained after the extraction of the a components

Finally, the multinomial model parameters can be expressed in function of original variables (**X** matrix).

$$Z^{(a)}\gamma^{(a)} = XV^{(a)}\gamma^{(a)} = X\beta^{(a)} \quad (6)$$

where, $\beta^{(a)} = V^{(a)}\gamma^{(a)}$; $Z^{(a)}$ is the matrix of a PC's; $\gamma^{(a)}$ is the matrix of parameters on a PC's for the s alternatives; $V^{(a)}$ is the matrix of a eigenvectors; $\beta^{(a)}$ is the matrix of parameters expressed in function of original variables. An interesting result which has been obtained is $\beta^{(p)} = \beta$, that is, if we retain all PC's in the model, the matrix of parameters expressed in function of original variables, $\beta^{(p)}$ is equal to the matrix of classical multinomial parameters, β.

However, the most important result is that the PCMR leads to lower variance estimates of model parameters comparing to classical multinomial model. We calculate the variance of the estimated parameters of the multinomial model by bootstrap resampling. Let, $\hat{\beta}_{jbl}^{(a)}$ be the bootstrap estimate of the parameter $\beta_{jb}^{(a)}$ for the *l*-th sample, let $\hat{\beta}_{jb}^{(a)}$ be the estimated parameter, the bootstrap estimate of variance of $\hat{\beta}_{jb}^{(a)}$ is the empirical estimate calculated from m bootstrap values:

$$S^2(\hat{\beta}_{jb}^{(a)}) = \frac{1}{m}\sum_{1}^{m}\left(\hat{\beta}_{jbl}^{(a)} - \bar{\beta}_{jb}^{(a)}\right)^2 \quad (7)$$

where, $\bar{\beta}_{jb}^{(a)} = \frac{1}{m}\sum_{1}^{m}\hat{\beta}_{jbl}^{(a)}$ is the bootstrap mean of the estimations of the j-th parameter

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 $1 \leq a \leq p$. 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:

- To consider the PC's in their natural order and stop when explained variability is about 75%
- 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 propose the criterion of considering in the model all the PC's that influence in statistical significant manner the response variable. A forward stepwise procedure is applied for selecting the significant components.

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 matrix. We propose two accuracy measures:

- The Root Mean Squared Error (RMSE) of bootstrap estimates for $\hat{\beta}_{jb}^{(a)}$
- The BIAS for $\hat{\beta}_{jb}^{(a)}$, calculated as the differences, in absolute value, between the bootstrap mean of the parameter estimations and the true values of the parameters

They are defined as follows:

$$\text{RMSE}\left(\hat{\beta}_{jb}^{(a)}\right) = \frac{1}{m}\sum_{1}^{m}\left(\hat{\beta}_{jbl}^{(a)} - \beta_{jb}\right)^2 \text{ and } \text{BIAS}\left(\hat{\beta}_{jb}^{(a)}\right) = \left|\bar{\beta}_{jb}^{(a)} - \beta_{jb}\right| \tag{8}$$

The simulation study and the accuracy measures show the best results are obtained, when the criterion of significant components (the third above-written-criterion) is used.

A SIMULATION STUDY

In order to illustrate the performance of the proposed approach, we develop a simulation study according to the scheme proposed by Hosmer *et al.* (1997).

The first step in the simulation process is to obtain a set of p explicative variables with a known correlation framework. For this purpose, we apply Cholesky decomposition. The second step is to fix a vector of real parameters β and compute the real probabilities. Finally, each value of the response is simulated from a multinomial distribution. After each data simulation, we fit the multinomial model. As we will see the estimated parameters $\hat{\beta}$ are always very different to the real ones due to multicollinearity.

As we stated in previous sections, the PCA of the regressors helps to improve this inaccurate estimation of the parameters. Once the PC's of the simulated covariates are computed, we fit the PCMR(a) models with different sets of a PC's. Then, we compute for all fitted PCMR(a) models the estimated parameters in terms of the original variables and their variance defined in the Eq. 7 for testing the improvement in the parameter estimation.

This simulation design is carried out for three different numbers of regressors (p = 10, 12 and 15) two different sample sizes (n = 100 and 200) three different number of alternatives (s = 3, 4 and 5) and two different distributions of regressors (standard normal and uniform distribution). The number of performed simulations is 360. Different criteria are considered to decide the number of components to retain in the model.

We present two specific simulation studies:

- Simulation 1 with n = 100, p = 10 and s = 4, correlation among the regressors from 0.4 to 0.9 and regressors with standard normal distribution
- Simulation 2 with n = 100, p = 10 and s = 4, correlation among the regressors from 0.4 to 0.9 and regressors with uniform distribution

Table 1 to 5 shows the results for the simulation 1 and their validation. Table 6 displays the results for the simulation 2.

Let us focus on the simulation 1. In the first column of table 1 there are the labels of the parameters, then we have the real parameters (Real), the parameters estimated with all the components (All PC's), the parameters calculated on the dataset of PC's with eigenvalue bigger than one (PCMR(1)) and the parameters estimated on the PC's that influence in significant manner the dependent variable (PCMR (sign)). In all the cases the parameters are calculated for the three possible alternatives (the fourth is the fixed alternative). As $\beta^{(p)} = \beta$, we do not insert the column of parameters obtained trough classical multinomial model.

It is easy to see (in bold character) that the parameters estimated with all the components have many discordances in the sign (15 discordant signs) and they are always very different

Table 1: Estimated parameters with the different methods for all the alternatives (simulation 1)

	Real			All PC's			PCMR(1)			PCMR (sign)		
Parameters	1	2	3	1	2	3	1	2	3	1	2	3
β_1	0.637	0.511	1.119	3.236	4.340	2.421	0.034	0.168	**-0.029**	**-0.754**	0.842	0.838
β_2	0.307	0.692	-1.665	**-16.488**	**-0.590**	-9.411	0.040	0.196	-0.034	0.211	0.820	-0.688
β_3	-1.121	1.790	-0.385	**14.263**	5.348	**6.295**	**0.037**	0.184	-0.032	-0.546	1.977	-0.354
β_4	-1.739	-0.590	0.175	**9.348**	-0.235	5.029	**0.041**	**0.201**	**-0.035**	**0.008**	-0.183	0.426
β_5	0.605	-0.593	-1.270	21.788	**4.932**	**9.293**	0.038	**0.190**	-0.033	0.497	**0.219**	-0.703
β_6	1.569	-0.720	0.379	**-3.847**	-3.393	**-3.115**	0.037	**0.182**	**-0.032**	0.714	-1.204	0.026
β_7	0.420	0.893	-0.024	**-20.006**	**-3.696**	-10.348	0.038	0.187	-0.033	**-0.462**	1.262	**0.062**
β_8	0.193	0.245	0.461	**-2.105**	0.015	0.860	0.038	0.186	**-0.032**	0.129	**-0.844**	0.715
β_9	-0.720	-0.291	0.770	-7.269	-0.696	**-3.169**	**0.039**	**0.192**	**-0.034**	**0.326**	**0.664**	**-0.763**
β_{10}	-0.848	1.506	-0.686	**1.519**	1.349	**1.890**	**0.041**	0.201	-0.035	-0.148	1.044	-0.264

Bold values indicate the parameters estimated with all components have many discordances in the sign

Table 2: Differences in absolute value between the real parameters and the estimated ones for all the alternatives (simulation 1)

	All PC's			PCMR(1)			PCMR (sign)		
Parameters	1	2	3	1	2	3	1	2	3
β_1	2.5990	3.8291	1.3016	0.6026	0.3424	1.1488	1.3903	0.3311	0.2816
β_2	16.7950	1.2824	7.7454	0.2673	0.4958	1.6311	0.0962	0.1283	0.9775
β_3	15.3842	3.5577	6.6800	1.1585	1.6056	0.3528	0.5747	0.1872	0.0310
β_4	11.0862	0.3551	4.8547	1.7794	0.7918	0.2098	1.7466	0.4080	0.2516
β_5	21.1833	5.5247	10.5628	0.5661	0.7824	1.2365	0.1079	0.8116	0.5666
β_6	5.4155	2.6735	3.4945	1.5322	0.9014	0.4111	0.8550	0.4845	0.3538
β_7	20.4252	4.5887	10.3247	0.3819	0.7062	0.0089	0.8817	0.3688	0.0861
β_8	2.2974	0.2298	0.3997	0.1553	0.0588	0.4931	0.0639	1.0886	0.2539
β_9	6.5496	0.4056	3.9396	0.7587	0.4831	0.8039	1.0460	0.9548	1.5331
β_{10}	2.3667	0.1568	2.5766	0.8883	1.3052	0.6511	0.6993	0.4622	0.4226

Table 3: RMSE of estimated parameters for the different methods for all the alternatives (simulation 1)

	All PC's			PCMR(1)			PCMR (sign)		
Parameters	1	2	3	1	2	3	1	2	3
β_1	129929.80	39891.33	148.7349	1.3690	0.1145	1.3409	1.8128	0.7109	0.2306
β_2	636260.70	292224.00	1041.5970	0.0762	0.2405	2.6276	0.1284	0.6317	0.8434
β_3	539104.60	238524.40	762.7269	1.3363	2.5556	0.1184	0.3864	2.6405	0.0892
β_4	192992.40	101793.30	423.5504	3.1555	0.6433	0.0495	2.0374	0.1604	0.1249
β_5	1307155.00	541290.70	1912.4260	0.3271	0.6272	1.5050	0.2048	1.0414	0.3182
β_6	49897.53	30350.83	110.6447	2.3604	0.8286	0.1777	0.7197	0.5582	0.2108
β_7	1056853.00	401339.50	1596.6380	0.1510	0.4903	0.0011	1.0303	1.3331	0.0487
β_8	28678.74	17138.12	44.0576	0.0276	0.0051	0.2536	0.1264	1.5096	0.2168
β_9	113074.60	50578.79	235.2796	0.5727	0.2439	0.6635	1.4803	1.8680	1.8031
β_{10}	12606.73	4923.52	14.3557	0.7852	1.6846	0.4110	0.4973	0.5432	0.1602

Table 4: BIAS: differences between the mean of the parameter estimations and the true values for all the alternatives (simulation 1)

	All PC's			PCMR(1)			PCMR (sign)		
Parameters	1	2	3	1	2	3	1	2	3
β_1	2.6853	40.6294	3.9948	0.6058	0.3354	1.1577	1.5730	0.5218	0.1836
β_2	102.6067	6.3430	14.8113	0.2710	0.4876	1.6207	0.0245	0.4512	0.8753
β_3	100.6865	70.2280	11.8090	1.1550	1.5979	0.3430	0.4837	0.8456	0.0310
β_4	50.9723	19.8728	9.0736	1.7756	0.8002	0.2205	1.7308	0.3123	0.2979
β_5	158.1853	72.2657	18.9201	0.5697	0.7903	1.2265	0.0203	0.9456	0.4649
β_6	15.6408	35.5992	5.5480	1.5356	0.9090	0.4207	0.7156	0.8670	0.3552
β_7	135.0317	41.7872	18.4592	0.3854	0.6984	0.0188	0.9739	0.7630	0.0807
β_8	21.1052	30.2797	0.2203	0.1588	0.0510	0.5029	0.0684	1.4197	0.3408
β_9	67.6659	32.2323	7.1343	0.7550	0.4912	0.8141	1.1435	1.2351	1.6442
β_{10}	27.6460	16.4636	3.1889	0.8845	1.2968	0.6404	0.6825	0.1072	0.3748

Table 5: Bootstrap estimates of variance of the parameters for the different methods for all the alternatives (simulation 1)

	All PC's			PCMR (sign)		
Parameters	1	2	3	1	2	3
β_1	131234.90	38626.85	134.1178	0.341915	0.442997	0.198862
β_2	632053.10	295135.10	830.5276	0.129102	0.432396	0.07801
β_3	534309.90	235952.00	629.5696	0.153923	1.944938	0.089099
β_4	192317.40	102422.60	344.6664	0.042144	0.063459	0.03653
β_5	1295083.00	541483.20	1570.157	0.206413	0.14868	0.103094
β_6	50154.44	29377.30	80.67084	0.209824	0.814682	0.085483
β_7	1049110.00	403629.70	1268.583	0.082631	0.758493	0.042659
β_8	28518.49	16385.11	44.45363	0.122988	0.498977	0.101682
β_9	109591.90	50040.27	186.2431	0.174383	0.345962	0.100577
β_{10}	11962.06	4699.461	4.228807	0.031753	0.537071	0.019927

Table 6: Estimated parameters with the different methods for all alternatives (simulation 2)

	Real			All PC's			PCMR(1)			PCMR (sign)		
Parameters	1	2	3	1	2	3	1	2	3	1	2	3
β_1	0.637	0.511	1.119	**-16.755**	**-3.106**	8.721	0.010	0.214	**-0.066**	**-3.932**	**-1.990**	3.567
β_2	0.307	0.692	-1.665	24.498	**-0.355**	-10.441	0.011	0.241	-0.074	0.638	2.232	-2.121
β_3	-1.121	1.790	-0.385	**3.477**	1.877	**6.596**	**0.009**	0.204	-0.063	**6.507**	0.626	-0.393
β_4	-1.739	-0.590	0.175	-15.951	**0.414**	5.264	**0.011**	**0.229**	**-0.071**	-12.90	-0.203	1.644
β_5	0.605	-0.593	-1.270	**-3.670**	-0.773	-1.653	0.009	**0.198**	-0.061	**-1.066**	-0.066	-1.600
β_6	1.569	-0.720	0.379	22.049	-3.727	**-2.306**	0.011	**0.233**	**-0.072**	6.821	-0.878	**-0.334**
β_7	0.420	0.893	-0.024	7.960	4.990	-0.879	0.009	0.207	-0.064	4.355	0.754	**0.046**
β_8	0.193	0.245	0.461	**-21.287**	3.355	5.122	0.011	0.248	**-0.076**	2.703	0.779	**-2.287**
β_9	-0.720	-0.291	0.770	**17.366**	**1.115**	**-12.813**	**0.011**	**0.248**	**-0.077**	-1.865	-0.275	1.302
β_{10}	-0.848	1.506	-0.686	-18.466	**-0.219**	**6.501**	**0.010**	0.220	-0.068	-1.875	1.719	**0.745**

Bold values indicate the parameters estimated with all components have many discordances in the sign

to the real ones. The situation improves if we consider the parameters calculated with the other two methods and we have the best results for the PCMR(sign): 9 signs discordant and parameters more similar to real ones.

For the sake of facilitating the reading we calculate in Table 2 the differences, in absolute value, between the real parameters and the estimated ones, with the different methods. It is possible observe that such differences are high for the classical multinomial method. As we have previously stated, these results must be caused by multicollinearity.

In order to validate the results obtained from each simulation we apply bootstrap resampling. Table 3 and 4 show the RMSE and the BIAS for the different methods, parameters and alternatives. The number of bootstrap samples is 100. In the Table 3, we can note that the RMSE, computed considering all the Principal Components, is very high for the first 2 alternatives, assuming, sometimes, values higher than 1000000. For the third alternative, the situation is a little better, but the results are not so good, if we compare them to the ones obtained with the techniques based on PCMR. In fact, form Table 3, that the values of the RMSE are always lower than before (not higher than 4) and, for the PCMR(sign), they are very low.

Passing to consider the Table 4 calculated BIAS, which is very high for the classical method. This means that the averages of parameter estimations are very far from the real values of the coefficients. If the reader looks at the results relative to the PCMR(1) and PCMR(sign), he can observe that they are better and, in particular, the values for the last method show that the bootstrap means of estimates are very near to the real parameters.

Finally in Table 5, we calculate the bootstrap estimates of variance of the estimated parameters using the Eq. 7. It is interesting to observe the variance of PCMR estimates is lower than classical multinomial one. In particular, the variance of classical multinomial estimates varies from 4.228807 to 1049110, PCMR (sign) one, instead, from 0.019927 to 1.944938. So, the aim of obtaining a considerable reduction in variability of estimates is reached. The positive effect of this result on confidence intervals, test hypothesis, etc., is remarkable.

Table 6 shows the results for the simulation 2. They confirm that the parameters estimated with all the components have many discordances in the sign (14 discordant signs) and they are always very different to the real ones.

The situation improves if we consider the parameters calculated with the other two methods and we have the best results for the PCMR(sign): 8 discordant signs and parameters more similar to real ones.

CONCLUSION

In this study, we showed the estimation of the multinomial logit model parameters presents a high variance, when there is multicollinearity among the regressors. To solve the problem, we proposed to use as covariates of the multinomial model a reduced number of PC's of the predictor variables. An extensive simulation study showed that the proposed approach is a valid alternative in presence of multicollinearity

In order to select the optimum PCMR(a) model, we considered and compared different methods for including PC's in the model. The method, that considered in model all the PC's that influence in statistical significant manner the response variable, yielded more stable and lower variance estimates of model parameters. This is a considerable advantage that will allow us to focus on inferential aspects of the proposed approach.

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

REFERENCES

Aguilera, A.M., M. Escabias and M.J. Valderrama, 2006. Using principal components for estimating logistic regression with high-dimensional multicollinear data. Comput. Stat. Data Anal., 50: 1905-1924.

Bastien, P., V. Esposito and M. Tenenhaus, 2005. PLS generalised linear regression. Comput. Stat. Data Anal., 48: 17-46.

Ben-Akiva, M. and S. Lerman, 1985. Discrete Choice Analysis. MIT Press, Cambridge, ISBN: 0262022176, pp: 390.

Frank, I.E., J.H. Friedman, S. Wold, T. Hastie and C. Mallows, 1993. A statistical view of some chemometrics regression tools. Technometrics, 35: 109-148.

Hoerl, A.E. and R.W. Kennard, 1970. Ridge regression: Biased estimation for nonorthogonal problems. Technometrics, 12: 55-67.

Hosmer, D.W., T. Hosmer, S. Le Cessie and S. Lemeshow, 1997. A comparison of goodness-of-fit tests for the logistic regression model. Stat. Med., 16: 965-980.

Luce, R.D., 1959. Individual Choice Behaviour: A Theoretical Analysis. John Wiley and Sons, New York.

Marschak, J., 1960. Binary Choice Constraints on Random Utility Indicators. In: Stanford Symposium on Mathematical Methods in the Social Scienses, Arrow, K. (Ed.). Stanford University Press, Stanford.

Marx, B.D., 1992. A continuum of principal component generalized linear regression. Comput. Stat. Data Anal., 13: 385-393.

Massy, W.F., 1965. Principal components regression in exploratory statistical research. J. Am. Stat. Assoc., 60: 234-256.

McFadden, D., 1974. Conditional Logit Analysis of Qualitative Choice Behavior. In: Frontiers in Econometrics, Zarembka, P. (Ed.). Academic Press, New York, pp: 105-142.

Ryan, T.P., 1997. Modern Regression Methods. John Wiley and Sons, Inc., New York.

Train, K.E., 2003. Discrete Choice Methods with Simulation. 1st Edn. Cambridge University Press, Cambridge.

Vágó, E. and S. Kemény, 2006. Logistic ridge regression for clinical data analysis (a case study). Applied Ecol. Environ. Res., 4: 171-179.

Wold, H., 1966. Estimation of Principal Components and Related Models by Iterative Least Squares. In: Multivariate Analysis, Krishnaiaah, P.R. (Eds.). Academic Press, New York, pp: 391-420.

Wold, H., 1985. Partial Least Squares. In: Encyclopedia of Statistical Sciences, Kotz, S. and N.L. Johnson (Eds.). Vol. 6, Wiles Publishers, New York, pp: 581-591.