Advances in Adaptive Data Analysis
Vol. 3, No. 3 (2011) 309–324

DOI: 10.1142/S1793536911000738

World Scientific
www.worldscientific.com

CONDITIONAL AND MULTINOMIAL LOGITS AS BINARY LOGIT REGRESSIONS

STAN LIPOVETSKY

GfK Custom Research North America,
8401 Golden Valley Road, Minneapolis, MN 55427, USA
stan.lipovetsky@gfk.com

For a categorical variable with several outcomes, its dependence on the predictors is usually considered in the conditional or multinomial logit models. This work considers elasticity features of the binary and categorical logits and introduces the coefficients individual by observations. The paper shows that by a special rearrangement of data the more complicated conditional and multinomial models can be reduced to binary logistic regression. It suggests the usage of any software widely available for logit modeling to facilitate constructing for complex conditional and multinomial regressions. In addition, for binary logit, it is possible to obtain meaningful coefficients of regression by transforming data to the linear link function, which opens a possibility to obtain meaningful parameters of the complicated models with categorical dependent variables.

Keywords: Binary conditional and multinomial logits; elasticity; multicollinearity.

1. Introduction

Discrete choice modeling by multiple predictors is widely used in regression analysis. Dichotomy response is often performed in the logistic approach [Long (1997); McCullagh and Nelder (1997); Ripley (1997); Lloid (1999); Lipovetsky (2006, 2008a, 2010a)]. Categorical variables with several outcomes have been developed in conditional and multinomial logits (MNLs) modeling and used in various applications [McFadden (1973, 1981, 1984); Hausman and McFadden (1984); McFadden and Richter (1990); Ben-Akiva and Lerman (1985); Arminger *et al.* (1995); Wedel and Kamakura (1999); Louviere *et al.* (2000); Hastie *et al.* (2001); Train (2003); Berry *et al.* (2004); Bishop (2006); Lipovetsky (2008b, 2009a, b, c); Greene and Hensher (2010)]. The current paper describes characteristics of elasticity for the models of binary logit (BL), conditional logit (CL), MNL, and a combined CL/MNL model. This consideration shows that there is a useful opportunity to rearrange data for more complicated models, so that any of them can be reduced to BL model. It suggests a possible usage of any software widely available for BL modeling to facilitate constructing of conditional and multinomial regressions.

Another important aspect of a wider application of BL model consists in the obtaining meaningful coefficients of regression for it. The problem is that for any objective for constructing a regression, is it the ordinary least squares for linear models, or maximum likelihood for nonlinear models, the numerical estimation of the parameters includes covariance or correlation matrix, or their analogue — a Hessian matrix of the second derivatives of the objective by the estimated parameters. The estimation of the parameters in linear modeling requires inverting covariance matrix to find the solution. In nonlinear estimation, a procedure of Newton–Raphson kind is applied for iterative solving that includes inversion of the Hessians. If such a matrix has highly correlated variables, its determinant is close to zero; therefore, inversion yields exceedingly inflated estimates covering a wide range of values of both signs, so more complicated models can easily produce pointless parameters. The model can have the best linear aggregate of the predictors to fit and predict data; however, it could be useless for analysis of the individual predictor impact on the response variable. For instance, if a presumably beneficial variable received a negative coefficient in BL model, should we use its higher or lower value to get a lift in the outcome?

The recent work [Lipovetsky (2010b)] considers how to reduce the effects of multicollinearity not only in linear functions but also in linear link functions, and to produce meaningful coefficients of regression. By transformation to BL model, the meaningful parameters of more complicated CL/MNL regressions can be attained. Consideration of the marginal effects, elasticity, and half-elasticity of the BL, CL, and MNL probabilities permits to find the adequate interpretation of the coefficients in each model, and even to define the coefficients of the individual observations.

The paper is organized as follows. Sections 2–5 consider some main features of binary, conditional, multinomial, and mixed CL/MNL logits, respectively, presenting them all in a binary response model, and Sec. 6 summarizes.

2. Binary Logit

Let us describe some main characteristics of the dichotomy output in the BL regression. A 0–1 variable p can be considered in two shares by the predictors:

$$p_{ik} = \frac{\exp(a_0^{(k)} + a_1^{(k)} x_{i1} + \cdots + a_n^{(k)} x_{in})}{\exp(a_0^{(1)} + a_1^{(1)} x_{i1} + \cdots + a_n^{(1)} x_{in}) + \exp(a_0^{(2)} + a_1^{(2)} x_{i1} + \cdots + a_n^{(2)} x_{in})}, \quad k = 1, 2. \tag{1}$$

Index i denotes observations $i = 1, 2, \ldots, N$ and x_{ij} is an ith observation by a jth predictor x_j, with $j = 1, 2, \ldots, n$, where N and n are the total numbers of observations and variables, respectively. The first share p_{i1} (for $k = 1$, with parameters $a_0^{(1)}, a_1^{(1)}, \ldots, a_n^{(1)}$) defines the theoretical model of probability of the outcome 0; therefore, the corresponding dummy vector d_{i1} of N binary observations contains ones on places when the event did not occur and zeros otherwise. The second

theoretical share p_{i2} (for $k = 2$, with parameters $a_0^{(2)}, a_1^{(2)}, \ldots, a_n^{(2)}$) defines probability of the outcome 1; therefore, the corresponding to it dummy vector d_{i2} of observations contains ones when the event did occur and zeros otherwise.

Both shares are complimentary to their total equal identically one, $p_{i1} + p_{i2} = 1$ for any observation; therefore, one share can be constructed in the model and the other is defined by this relation. Let us consider the share p_{i2} of the event occurrence. As a quotient, it is homogeneous in multiplying the numerator and denominator by any nonzero term, or by adding an item into the exponents. Therefore, the share p_{i2} (1) can be divided in its numerator and denominator by the complementary exponent of $k = 1$:

$$p_{i2} = \frac{\exp((a_0^{(2)} - a_0^{(1)}) + (a_1^{(2)} - a_1^{(1)})x_{i1} + \cdots + (a_n^{(2)} - a_n^{(1)})x_{in})}{1 + \exp((a_0^{(2)} - a_0^{(1)}) + (a_1^{(2)} - a_1^{(1)})x_{i1} + \cdots + (a_n^{(2)} - a_n^{(1)})x_{in})}. \tag{2}$$

Therefore, the probability of the event actually depends on the differences of the coefficients of two sets used in Eq. (1). Denoting these differences as new parameters, $a_j = a_j^{(2)} - a_j^{(1)}$, and the probability of the event occurrence simply as $p_i \equiv p_{i2}$, the expression (2) can be represented as following:

$$p = \frac{\exp(a_0 + a_1 x_1 + \cdots + a_n x_n)}{1 + \exp(a_0 + a_1 x_1 + \cdots + a_n x_n)}. \tag{3}$$

It can be obtained directly from the second share (1) if to choose the first set of parameters as the reference, so to put them to zero:

$$a_0^{(1)} = a_1^{(1)} = \cdots = a_n^{(1)} \equiv 0. \tag{4}$$

Formula (3) shows a common form of the logit model, written for simplicity without the index i of observations in the predictors x_j and the outcome p.

After estimation of the coefficients in Eq. (3), the prediction of the values p for each observation can be obtained. The predicted values present a continuous variable in the range $p \in [0,\ 1]$, which estimates the binary event probability for each observation point or for new values of the predictors. By the predicted values p, the model (3) can be represented in a linear link function:

$$y \equiv \ln \frac{p}{1-p} = a_0 + a_1 x_1 + a_2 x_2 + \cdots + a_n x_n, \tag{5}$$

where y denotes the dependent variable defined by logarithm of probability odds. Partial derivative of Eq. (3) by a predictor x_j gives its marginal effect in change of probability:

$$\frac{\partial p}{\partial x_j} = a_j p(1-p), \tag{6}$$

where p in Eq. (6) is defined in each ith observation by the probability (3). While the expression (3) corresponds to the cumulative probability, the product $p(1-p)$ in Eq. (6) defines the probability density function for the logit distribution.

Division of the relation (6) by p yields a half-elasticity, or the percentage change in probability due to an absolute change in the predictor:

$$\frac{\frac{\partial p}{p}}{\partial x_j} = a_j(1-p), \tag{7}$$

where p depends on an ith observation. Dividing Eq. (7) by $1-p$ produces an expression useful in the interpretation of the coefficient of the logit model:

$$\begin{aligned} a_j &= \frac{\partial p}{\partial x_j}\frac{1}{p(1-p)} = \frac{\partial p}{\partial x_j}\left(\frac{1}{p}+\frac{1}{1-p}\right) \\ &= \frac{\partial \ln p}{\partial x_j} - \frac{\partial \ln(1-p)}{\partial x_j} = \frac{\partial}{\partial x_j}\ln\frac{p}{1-p}. \end{aligned} \tag{8}$$

It shows that a change in logarithm of the probability odds due to a predictor x_j change equals the constant value of the coefficient a_j of the logit model. This result can be easier obtained from the linear link (5), and simplified to the expression:

$$a_j = \frac{\partial y}{\partial x_j} = \frac{\partial}{\partial x_j}\ln\frac{p}{1-p} = \frac{\partial\left(\frac{p}{(1-p)}\right)}{\partial x_j}\Bigg/\left(\frac{p}{(1-p)}\right). \tag{9}$$

Thus, a coefficient of BL model makes sense of the relative change in the probability's odds due to a change in the corresponding predictor when all the other predictors are held fixed.

Although the expressions (3)–(9) describe the logit model in the simple forms, the expressions (1) and (2) are more convenient for better interpretation and comparison with complicated multinomial models; therefore, they are useful as well. For instance, to understand the meaning of the relation (7) and to get some results from it, consider a partial derivative of the kth share (1) by a predictor x_j, so its marginal effect in an ith point of observation:

$$\frac{\partial p_{ik}}{\partial x_{ij}} = p_{ik}a_j^{(k)} - p_{ik}(p_{i1}a_j^{(1)} + p_{i2}a_j^{(2)}) = p_{ik}(a_j^{(k)} - \bar{a}_{ij}), \tag{10}$$

where the bar denotes the mean of the coefficients for each predictor x_j weighted by the shares of probability (1):

$$\bar{a}_{ij} = p_{i1}a_j^{(1)} + p_{i2}a_j^{(2)}. \tag{11}$$

Dividing in Eq. (10) by p_{ik} yields the following expression for the half-elasticity of the relative change in probability of an outcome due to the absolute change in one predictor only:

$$\frac{\frac{\partial p_{ik}}{p_{ik}}}{\partial x_{ij}} = a_j^{(k)} - \bar{a}_{ij} \equiv b_{ij}^{(k)}, \tag{12}$$

which equals the deviation $b_{ij}^{(k)}$ of the kth share's coefficients from the mean (11).

With the setup (4) and notations $p_i \equiv p_{i2}$ and $a_j = a_j^{(2)}$, the mean (11) can be represented as $\bar{a}_{ij} = p_i a_j$. Using this mean in Eq. (10) yields the same expression (6), and using it in Eq. (12) yields the expression:

$$b_{ij}^{(1)} = -p_i a_j, \quad b_{ij}^{(2)} = a_j - p_i a_j. \tag{13}$$

The second share's coefficients coincide with those given in Eq. (7). It suggests a very interesting explanation for the meaning of the expressions (7) and (12): the coefficients $b_{ij}^{(k)}$ in Eqs. (12) and (13) can be interpreted as the individual coefficients of the logit model for each ith observation. Then, the half-elasticities (7) and (12) equal the individual coefficients of the logit regression. It is clear from Eqs. (12) and (13) that the mean $\bar{b}_{ij} = p_{i1} b_{ij}^{(1)} + p_{i2} b_j^{(2)} = 0$; therefore, the new coefficients $b_{ij}^{(k)}$ are standardized by this condition, not by the condition (4).

Using these varying across the observations coefficients $b_{ij}^{(k)}$ in place of the previous constant by observations parameters $a_j^{(k)}$ does not change the probabilities in the shares (1). Indeed, in any ith observation, the formula (1) can be reduced to the expression (2) that depends on the coefficients' difference $a_j = a_j^{(2)} - a_j^{(1)}$. If to substitute $b_{ij}^{(k)}$ instead of $a_j^{(k)}$ into the formula (1) and transform it to Eq. (2), the last expression would be represented via the differences of the new coefficients, so:

$$b_{ij}^{(2)} - b_{ij}^{(1)} = (a_j^{(2)} - \bar{a}_j) - (a_j^{(1)} - \bar{a}_j) = a_j, \tag{14}$$

which are the same parameters of the logit model (2). Thus, using the individual coefficients (13) does not change the shares' probability (1) and (2) because those are defined only by the differences of the coefficients (14), which do not depend on observation. However, if needed, for the utility evaluations as the inputs from $b_{ij}^{(k)} x_{ij}$ in the total score y_i (5), the individual coefficients for each observation can be easily found by the constructed logit coefficients a_j and the predicted with them values of the probability p_i (13).

3. Conditional Logit

Let us consider a categorical dependent variable of several (K) outcomes, or choices. Extension of a binary to a multiple outcome can be presented in multinomial-logit models. For a multiple categorical outcome, the two models are usually applied — those are the so-called CL and MNL. They differ in the specifics of the predictors' relation to the choice.

If the values of the predictors $x_j^{(k)}$ are different for each kth choice, then it is a CL model widely known in conjoint and discrete choice modeling in economics and marketing research. It can be specified as:

$$p_{ik} = \frac{\exp(a_0 + a_1 x_{i1}^{(k)} + \cdots + a_n x_{in}^{(k)})}{\sum_{q=1}^{K} \exp(a_0 + a_1 x_{i1}^{(q)} + \cdots + a_n x_{in}^{(q)})}, \tag{15}$$

where the set of parameters $a_0, a_1, \ldots, a_n$ is the same for all the shares $k, q = 1, 2, \ldots, K$, and the total of the shares equals 1 for each ith observation:

$$p_{i1} + p_{i2} + \cdots + p_{iK} = 1. \tag{16}$$

A partial derivative of a share (15) by a predictor is given by the expression:

$$\frac{\partial p_{ik}}{\partial x_{ij}^{(q)}} = a_j p_{ik}(\delta_{kq} - p_{iq}), \tag{17}$$

where δ_{kq} is Kronecker's delta, which equals one or zero when $q = k$ or $q \neq k$.

For the self-marginal effects ($q = k$), dividing the expression (17) by p_{ik} produces the same expression as Eq. (13) for each share:

$$\frac{\frac{\partial p_{ik}}{p_{ik}}}{\partial x_{ij}^{(q)}} = a_j(1 - p_{ik}) \equiv b_{ij}^{(k)}, \tag{18}$$

which corresponds to the BL expressions (7) and (13) for CL coefficient deviations for a kth share. Then, in the same interpretation, as in Eq. (13), the coefficients $b_{ij}^{(k)}$ can be used as the individual coefficients of the CL for each ith observation in the utility evaluations by the inputs from $b_{ij}^{(k)} x_{ij}$ in their total score.

In addition, for the self-marginal effects, the expression (17) can be represented for the coefficient a_j as:

$$a_j = \frac{1}{p_{ik}(1 - p_{ik})} \frac{\partial p_{ik}}{\partial x_{ij}^{(k)}} = \frac{\partial}{\partial x_{ij}^{(k)}} \ln \frac{p_{ik}}{1 - p_{ik}} = \frac{\partial \left(\frac{p_{ik}}{(1 - p_{ik})} \right)}{\partial x_{ij}^{(k)}} \Bigg/ \left(\frac{p_{ik}}{(1 - p_{ik})} \right). \tag{19}$$

Therefore, each coefficient a_j of CL model (15) is defined by the relative change of the probability's odds in any share p_k due to the change in a jth predictor $x_j^{(k)}$ and fixed other predictors. Thus, for any kth outcome of the CL (15), the same property (9) as for binary model (1) is held. It is a very useful property for interpretation of the CL coefficients as gradients of the relative odds for any share by its predictors.

The similarity of CL (15) to BL model (1) can be also seen if to take one share, say, the first one as the reference; therefore, to divide by the first exponent in Eq. (15), which yields:

$$p_{ik} = \frac{\exp(a_1(x_{i1}^{(k)} - x_{i1}^{(1)}) + \cdots + a_n(x_{in}^{(k)} - x_{in}^{(1)}))}{1 + \sum_{q=2}^{K} \exp(a_1(x_{i1}^{(q)} - x_{i1}^{(1)}) + \cdots + a_n(x_{in}^{(q)} - x_{in}^{(1)}))}. \tag{20}$$

This expression shows that CL model is essentially defined by the differences of the variables from those of a share taken as the reference. Then, shares without the first one can be used for the estimation of the regression coefficients and the complimentary share can be simply found by the relation (16). Different intercepts can be used in each share (15), and it would correspond to the additional items of intercept in Eq. (20) as well.

The property (19) for CL model can be proved in a straightforward taking of the complimentary for Eq. (15) of "not occurred" event with probability:

$$1 - p_{ik} = \frac{\sum_{q \neq k}^{K} \exp(a_0 + a_1 x_{i1}^{(q)} + \cdots + a_n x_{in}^{(q)})}{\sum_{q=1}^{K} \exp(a_0 + a_1 x_{i1}^{(q)} + \cdots + a_n x_{in}^{(q)})}. \tag{21}$$

Then, logarithm of probability's odds of Eqs. (15)–(21) equals:

$$\ln \frac{p_{ik}}{1 - p_{ik}} = a_0 + a_1 x_{i1}^{(k)} + \cdots + a_n x_{in}^{(k)} - \ln \left(\sum_{q \neq k}^{K} \exp(a_0 + a_1 x_{i1}^{(q)} + \cdots + a_n x_{in}^{(q)}) \right), \tag{22}$$

which is a link like Eq. (5) of the predictors in kth share. The last expression of logarithm of sum in Eq. (22) does not depend on the kth predictors at all. Therefore, a derivative of the logarithm of the odds ratio (22) by any predictor yields similar to Eq. (9) expression for the coefficient of regression (19) as the relative change in odds due to a predictor's change.

The identity of the expression (19) for a coefficient a_j by any share p_k indicates to a valuable possibility to estimate parameters of CL model by the BL for the combined data stacked in rows by all the shares. The output variable can be presented as a binary dummy D of all the stacked vectors d_k of the dummy binary values for all shares. Each dummy vector d_{ik} of N binary observations contains ones on places when the corresponding kth outcome occurs and zeros otherwise. The total matrix X_{total} of predictors is arranged as stacked in rows all K matrices $X^{(k)}$ of the predictors for each share (each matrix of the order N by $1 + n$, and has the first identity column corresponding to using of the intercept in the predictors aggregate). Together with the vector of $1 + n$ parameter, the main structures can be presented in matrix form as:

$$D = \begin{pmatrix} d_1 \\ d_2 \\ \cdots \\ d_K \end{pmatrix}, \quad X_{\text{total}} = \begin{pmatrix} X^{(1)} \\ X^{(2)} \\ \cdots \\ X^{(K)} \end{pmatrix}, \quad a = \begin{pmatrix} a_0 \\ a_1 \\ \cdots \\ a_n \end{pmatrix}. \tag{23}$$

For N observations, the combined vector D is of the NK order and the total matrix X_{total} has the same number of rows and $1 + n$ column of all variables. It is also possible to use a stacked binary outcome without one share, and the related total predictor matrix of the stacked differences like in Eq. (20) of the variables from a share used as the reference. Thus, the construction for CL model (15) can be done with the data (23) as a regular BL model with all its properties (1)–(14). Particularly, the coefficients of CL model (15) have the meaning of the relative change in the probability's odds (9) of all the stacked shares (23).

4. Multinomial Logit

When the outcomes are defined by individual observations of all the predictors, not by different variables related to each choice, it is the general MNL, which can be presented in the following model:

$$p_{ik} = \frac{\exp(a_0^{(k)} + a_1^{(k)} x_{i1} + \cdots + a_n^{(k)} x_{in})}{\sum_{q=1}^{K} \exp(a_0^{(q)} + a_1^{(q)} x_{i1} + \cdots + a_n^{(q)} x_{in})}, \quad k = 1, 2, \ldots, K, \tag{24}$$

where the sets of parameters $a_0^{(k)}, a_1^{(k)}, \ldots, a_n^{(k)}$ are different for each share $k, q = 1, \ldots, K$. MNL model (24) is a straightforward generalization of BL model (1) for more than two outcomes in the categorical dependent variable. For the sake of the parameters' identification, the first choice can be taken as the reference, and similarly to the BL model (2) and (3) the expression (24) can be represented as follows:

$$p_{ik} = \frac{\exp(a_0^{(k)} + a_1^{(k)} x_{i1} + \cdots + a_n^{(k)} x_{in})}{1 + \sum_{q=2}^{K} \exp(a_0^{(q)} + a_1^{(q)} x_{i1} + \cdots + a_n^{(q)} x_{in})}, \tag{25}$$

with zero coefficients (4) for the first share, which can be also found by the identity (16).

A partial derivative of a share (25) by a predictor can be presented as:

$$\frac{\partial p_{ik}}{\partial x_{ij}} = p_{ik} \left(a_j^{(k)} - \sum_{q=1}^{K} a_j^{(q)} p_{iq} \right) = p_{ik} (a_j^{(k)} - \bar{a}_{ij}), \tag{26}$$

where $\bar{a}_{ij}$ is the mean all the jth coefficients $a_j^{(q)}$ weighted by the choice probabilities (25) for the ith observation:

$$\bar{a}_{ij} = \sum_{q=1}^{K} a_j^{(q)} p_{iq}. \tag{27}$$

Then, the relative change in probability of a kth choice due to an absolute change in a jth variable equals the deviation $b_{ij}^{(k)}$ of each coefficient from its jth mean in an ith point:

$$\frac{\frac{\partial p_{ik}}{p_{ik}}}{\partial x_{ij}} = a_j^{(k)} - \bar{a}_{ij} \equiv b_{ij}^{(k)}. \tag{28}$$

The expressions (26)–(28) generalize those of Eqs. (10)–(12) from a binary to a multiple outcome. If to use the individual parameters $b_{ij}^{(k)}$ in place of the constant parameters $a_j^{(k)}$, the expression (28) becomes:

$$\begin{aligned} \frac{\frac{\partial p_{ik}}{p_{ik}}}{\partial x_{ij}} &= b_{ij}^{(k)} - \sum_{q=1}^{K} b_{ij}^{(q)} p_{iq} = \left(a_j^{(k)} - \sum_{q=1}^{K} a_j^{(q)} p_{iq} \right) - \sum_{g=1}^{K} \left(a_j^{(g)} - \sum_{q=1}^{K} a_j^{(q)} p_{iq} \right) p_{ig} \\ &= a_j^{(k)} - \sum_{q=1}^{K} a_j^{(q)} p_{iq} - \sum_{g=1}^{K} a_j^{(g)} p_{ig} + \sum_{q=1}^{K} a_j^{(q)} p_{iq} \sum_{g=1}^{K} p_{ig} = a_j^{(k)} - \sum_{q=1}^{K} a_j^{(q)} p_{iq} \end{aligned}, \tag{29}$$

where (16) is taken into account. Thus, the relation (28) is held in Eq. (29) where the individual parameters used. The final expression in Eq. (29) due to Eq. (28) equals the same coefficient $b_{ij}^{(k)}$, and their weighted total equals zero, that can be also seen from Eq. (28), so:

$$\bar{b}_{ij} = \sum_{k=1}^{K} b_{ij}^{(k)} p_{ik} = 0. \tag{30}$$

Thus, similarly to the meaning of the coefficients in BL model (12)–(14), the half-elasticity (28) defines varying by observation individual coefficients in MNL model, and Eq. (30) is the condition for their normalization. Substituting the coefficients $b_{ij}^{(k)}$ in place of $a_j^{(k)}$ does not change the probabilities and can be used for the utility evaluations as the inputs from $b_{ij}^{(k)} x_{ij}$ in their total aggregates in the shares (25).

A generalization of the expressions (8) and (9) for the multiple outcomes can be obtained by the following transformation of the formula (26):

$$\begin{aligned} \frac{\partial p_{ik}}{\partial x_{ij}} &= p_{ik}\left(a_j^{(k)} - \sum_{q=1}^{K} a_j^{(q)} p_{iq}\right) \\ &= p_{ik}\left(a_j^{(k)}(1-p_{ik}) - \sum_{q\neq k}^{K} a_j^{(q)} p_{iq}\right) = p_{ik}(1-p_{ik})(a_j^{(k)} - \tilde{a}_{ij}). \end{aligned} \tag{31}$$

The tilde in Eq. (31) marks the mean value of the coefficients without the kth share weighted by all the other shares:

$$\tilde{a}_{ij} = \frac{\sum_{q\neq k}^{K} a_j^{(q)} p_{iq}}{1-p_{ik}} = \frac{\sum_{q\neq k}^{K} a_j^{(q)} p_{iq}}{\sum_{q\neq k}^{K} p_{iq}} = \sum_{q\neq k}^{K} a_j^{(q)} \frac{p_{iq}}{\sum_{g\neq k}^{K} p_{ig}} = \sum_{q\neq k}^{K} a_j^{(q)} \gamma_{iq}, \tag{32}$$

where the relation (16) is used, and γ_{iq} denotes the new renormalized probabilities also summing to one. The weighted partial mean (32) differs from the total mean (27) by using the renormalized weights of probabilities without the kth share. In the same way, as Eq. (6) for BL or Eq. (17) for CL, the change in MNL share (31) contains the probability density function $p_{ik}(1-p_{ik})$, so dividing by it yields:

$$\begin{aligned} \frac{1}{p_{ik}(1-p_{ik})}\frac{\partial p_{ik}}{\partial x_{ij}} &= \frac{\partial}{\partial x_{ij}} \ln \frac{p_{ik}}{1-p_{ik}} \\ &= \frac{\frac{\partial\left(\frac{p_{ik}}{(1-p_{ik})}\right)}{\partial x_{ij}}}{\left(\frac{p_{ik}}{(1-p_{ik})}\right)} = a_j^{(k)} - \tilde{a}_{ij} \equiv c_{ij}^{(k)}, \end{aligned} \tag{33}$$

where $c_{ij}^{(k)}$ denotes the coefficients' deviations from the partial means. In contrast to BL (9) and CL (19), the relative change in the probability's odds ratio of MNL (33) equals not the coefficient itself but its deviation from the corresponding partial mean.

The property (33) can be obtained likewise in the derivation (21) and (22) taking the complimentary for Eq. (24) "not occurred" event with probability $1 - p_{ik}$, so the logarithm of probability's odds for MNL equals:

$$\ln \frac{p_{ik}}{1-p_{ik}} = a_0^{(k)} + a_1^{(k)} x_{i1} + \cdots + a_n^{(k)} x_{in} - \ln \left(\sum_{q \neq k}^{K} \exp(a_0^{(q)} + a_1^{(q)} x_{i1} + \cdots + a_n^{(q)} x_{in}) \right), \tag{34}$$

which is a link like (22) of the predictors in kth share. Taking a derivative of this logarithm of the odds ratio (34) by any predictor yields the expression (33) of a coefficient's deviation as the relative change in odds due to the absolute predictor's change.

Similarly to Eq. (30), the total of the deviations (33) from the means (32) equals zero too:

$$\bar{c}_{ij} = \sum_{k=1}^{K} c_{ij}^{(k)} \gamma_{ik} = 0, \tag{35}$$

because for the renormalized probabilities γ_{iq}, the relation (16) is also held. Due to Eq. (33), the individual coefficients $c_{ij}^{(k)}$ can be used in MNL model and Eq. (35) is their condition of normalization. Substituting the coefficients $c_{ij}^{(k)}$ in place of $a_j^{(k)}$ in the model (24) does not change the shares' probabilities and can be used for the utility inputs from $c_{ij}^{(k)} x_{ij}$ in their total aggregates in the shares. Thus, the individual parameters $b_{ij}^{(k)}$ or $c_{ij}^{(k)}$ would yield the same probability values in the shares (24), but they have different meaning of half-elasticity (28) or the relative change in the probability's odds ratio (33), respectively.

The MNL (24) can be reduced to the CL (15) in such a combining transformation. Denoting a vector-column of a kth share's coefficients (24) as $a^{(k)} = (a_0^{(k)}, a_1^{(k)}, \ldots, a_n^{(k)})'$ (with prime marking transposition) let us stack all the K shares' vectors into one vector-column A of the order $K(1+n)$. Let us also take the matrix X of the predictors in Eq. (24) (it is of N by $1+n$ order, where N is the number of observations and one identity column is added to the n predictor columns for accounting of the intercept), and a zero-matrix $0 \cdot X$ of the same order. Stacking in columns $K-1$ zero-matrices and one matrix X, where X is inserted on each kth place, we can call such K matrices as $X^{(k)}$. Then, any kth share of MNL (24) can be expressed via the following components:

$$d_k, \quad X^{(k)} = (0, \ldots, X, \ldots, 0), \quad A = (a^{(1)'}, \ldots, a^{(k)'}, \ldots, a^{(K)'})'. \tag{36}$$

The output variable d_k for each share is a dummy containing N binary observations with ones on places when the corresponding kth outcome occurs and zeros otherwise. The product of each matrix $X^{(k)}$ by the same total vector-column A (36) yields the aggregate of the predictors used in the kth share of MNL model (24). It is also possible to use only $K-1$ stacked matrices and vectors in Eq. (36)

taking only those corresponding to the reduced MNL shares (25). A model with one and the same vector of parameters used with different matrices of predictors corresponding to several outputs of a categorical variable is actually the CL model (15), thus, the transformation (36) presents the MNL model in the CL form.

If the multinomial model can be presented as the CL model (15), then all CL properties (16)–(23) can be used for MNL model as well. Particularly, the transformation (23) for presenting CL as the BL model can be also used for presenting MNL as BL model. Indeed, combining, as in Eq. (23), all the matrices $X^{(k)}$ and all the binary outputs d_k from Eq. (36) into one total matrix X_{total} and one vector D represents (36) as follows:

$$D = \begin{pmatrix} d_1 \\ d_2 \\ \cdots \\ d_K \end{pmatrix}, \quad X_{\text{total}} = \begin{pmatrix} X & 0 & \cdots & 0 \\ 0 & X & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \ldots & X \end{pmatrix}, \quad A = \begin{pmatrix} a^{(1)} \\ a^{(2)} \\ \cdots \\ a^{(K)} \end{pmatrix}. \tag{37}$$

The total block-diagonal matrix with the repeated data matrix X on diagonal and zero nondiagonal blocks, being multiplied by the stacked vector A of all the shares' coefficients yields the aggregates used in the MNL model (24). At the same time, the formulation (37) with the one combined vector D of all binary outputs presents the regular BL model (1).

5. Mixed Conditional–Multinomial Logit

There could be various mixed conditional–multinomial logit models of combined Eqs. (15) and (25) inputs to the aggregate of the predictors. Using the above-described transformations, these models can be expressed as the BL as well. Let us consider a more complicated case of a CL/MNL model with different predictors $x_j^{(k)}$ for each kth choice defined by different sets of the coefficients $a_j^{(k)}$ in the shares of probabilities specified as following:

$$p_{ik} = \frac{\exp(a_0^{(k)} + a_1^{(k)} x_{i1}^{(k)} + \cdots + a_n^{(k)} x_{in}^{(k)})}{\sum_{q=1}^{K} \exp(a_0^{(q)} + a_1^{(q)} x_{i1}^{(q)} + \cdots + a_n^{(q)} x_{in}^{(q)})}. \tag{38}$$

A partial derivative of a share (38) by a predictor is similar to CL (17) expression:

$$\frac{\partial p_{ik}}{\partial x_{ij}^{(q)}} = a_j^{(q)} p_{ik}(\delta_{kq} - p_{iq}), \tag{39}$$

but now there is a different coefficient $a_j^{(q)}$ for each share. For the self-marginal effects, dividing by p_{ik} yields:

$$\frac{\frac{\partial p_{ik}}{p_{ik}}}{\partial x_{ij}^{(q)}} = a_j^{(k)}(1 - p_{ik}) \equiv b_{ij}^{(k)}, \tag{40}$$

which in contrast to Eq. (18) is different for each share. The expression (40) corresponds to the BL expressions (7) and (13) for the coefficient deviations for a kth share. By the same interpretation as in Eq. (13), the individual coefficients $b_{ij}^{(k)}$ (40) can be used for each ith observation in the utility evaluations by the inputs from $b_{ij}^{(k)} x_{ij}$ in their total score.

The expression (39) for the self-marginal effects can also be represented as in Eq. (19):

$$a_j^{(k)} = \frac{\frac{\partial\left(\frac{p_{ik}}{(1-p_{ik})}\right)}{\partial x_{ij}^{(k)}}}{\left(\frac{p_{ik}}{(1-p_{ik})}\right)}. \tag{41}$$

Therefore, each coefficient $a_j^{(k)}$ of CL/MNL model (38) has a meaning of the relative change of the probability's odds ratio in the related share p_k due to the change in a jth predictor $x_j^{(k)}$ and fixed other predictors. The property (41) can be derived straightforwardly as in Eqs. (21) and (22). In contrast to the simplified presentations (2), (20), or (25) in the previous models, the mix model (38) cannot be expressed via differences in the sets of coefficients or in the data segments of the shares.

The same well as CL (23) and MNL (37), the mix model (38) can be also reduced to BL model by the transformation combining all the matrices $X^{(k)}$ and all the binary outputs d_k into one total matrix X_{total} and one vector D as follows:

$$D = \begin{pmatrix} d_1 \\ d_2 \\ \cdots \\ d_K \end{pmatrix}, \quad X_{\text{total}} = \begin{pmatrix} X^{(1)} & 0 & \cdots & 0 \\ 0 & X^{(2)} & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & X^{(K)} \end{pmatrix}, \quad A = \begin{pmatrix} a^{(1)} \\ a^{(2)} \\ \cdots \\ a^{(K)} \end{pmatrix}. \tag{42}$$

In contrast to the case of MNL (37), the total block-diagonal matrix in Eq. (42) consists of different data matrices which being multiplied by the stacked vector A of all the shares' coefficients yields the aggregates used in the combined Eq. (38). In addition, the formulation (42) also presents the regular BL model (1); therefore, construction of the model (38) can be done as a logit model by all the data stacked. On the other hand, separable block-diagonal matrices like in Eq. (42) indicate that it is possible to estimate parameters within each block also independently by logit models related to the data of each choice. Separate logits consider the parameters of each choice (3) as the event occurrence against all the other choices as its absence, while a multinomial model uses all the choices' parameters (38) explicitly. The intercept in the data corresponds to the option of no choice taken in each particular response.

Estimations of the parameters for the above-described BL (3), CL (15), MNL (25), or mix CL/MNL (38) models are usually performed by the maximum log-likelihood objective:

$$\log L = \sum_{i=1}^{N} \sum_{k=1}^{K} d_{ik} \ln p_{ik}, \tag{43}$$

where d_{ik} and p_{ik} correspond to the observed binary outcomes and their theoretical probability models, respectively. Due to the combined data presentations in Eqs. (23), (37), and (42), all the CL, MNL, and mix CL/MNL models can be reduced to BL model featured in Eqs. (1)–(14). It means that any of these complicated models can be constructed via much more easily available software for BL modeling. Interpretation of the coefficients in the complicated models can be given via the coefficient of BL model by the combined data as the relative change in the probability's odds ratio (9) due to a change in the corresponding predictor when all the other predictors are held fixed.

Different models considered above are often used by the generic name of multinomial model. In the original work by Daniel McFadden (1973), the choice models had been introduced as the CL. In his Nobel speech, "Economic Choices" in 2000, McFadden said about his utility models: "I called this a *conditional logit model* since in the case of binomial choice it reduced to the logistic model used in biostatistics, and in the multinomial case it could be interpreted as the conditional distribution of demand given the feasible set of choice alternatives C. Today, (Eq. (1)) is more commonly called the *multinomial logit* (MNL) model, and I will use this more common terminology." Most of the works neither specify which exactly model is used nor use the term conditional for the related models (for instance, McCullagh and Nelder (1997) describe the CL but by general name of multinomial model). Of course, it is not so important if the properties of each of those models are given to identify its specific form.

Concerning a question which model to prefer in which case, we can conclude that if a choice depends on the values of the predictors (not on individual, for instance, demographical characteristics) then the CL is the adequate model. Indeed, the logarithms of probabilities of two shares in CL (15) are related as:

$$\frac{\ln p_{ik}}{\ln p_{ij}} = a_1(x_{i1}^{(k)} - x_{i1}^{(j)}) + \cdots + a_n(x_{in}^{(k)} - x_{in}^{(j)}). \tag{44}$$

The individual characteristics (demographical, social, financial, and others) are the same for any kth and jth different options; therefore, the choices cannot depend on them in the CL model. Only the predictors with the values related to different choices play role in resulting probabilities of the preference (44). On the other hand, if the model should take into account specifics of the individual observations by the respondents, then MNL is the correct tool. Taking the logarithms of probabilities by two options of MNL (24), we have:

$$\frac{\ln p_{ik}}{\ln p_{ij}} = (a_0^{(k)} - a_0^{(j)}) + (a_1^{(k)} - a_1^{(j)})x_{i1} + \cdots + (a_n^{(k)} - a_n^{(j)})x_{in}. \tag{45}$$

We see that MNL model works for distinguishing of the choice probabilities (45) for any given value x_{ij} of each predictor. The mix CL/MNL model with different values of predictors $x_j^{(k)}$ for each kth choice and different individual influences on the choices defined by altered sets of the coefficients $a_j^{(k)}$ in the shares of probabilities

(38) yields the following quotient of logarithms of choices:

$$\frac{\ln p_{ik}}{\ln p_{ij}} = (a_0^{(k)} - a_0^{(j)}) + (a_1^{(k)} x_{i1}^{(k)} - a_1^{(j)} x_{i1}^{(j)}) + \cdots + (a_n^{(k)} x_{in}^{(k)} - a_n^{(j)} x_{in}^{(j)}). \quad (46)$$

This estimate of the choice preferences depends simultaneously on both individual and option features in a more complicated way. Of course, the appropriate model should be taken for each data due to its characteristics and the aims of the research.

Similar to the BL model, the CL has a set of the same $n+1$ parameters (including the intercept) for all the shares of probabilities, while the multinomial model uses $(n+1)(K-1)$ parameters for its $K-1$ independent shares, and the mix conditional-multinomial has even $(n+1)K$ parameters for its all K-independent shares. If possible, it is always preferred to use a simpler conditional model with minimum number of parameters; therefore, their numerical estimation in a nonlinear procedure would produce more robust results. The so-called Occam's razor principle (by the name of medieval philosopher William Occam) of using only the main necessary number of parameters can be applied for the choice of the models [see in Gregory (2010)]. For estimating parameters of CL, MNL, and mix CL/MNL model by the simple BL model, we can use the combined data stacked in rows by all the shares in Eqs. (23), (37), and (42), respectively. To construct meaningful parameters of the models, a detailed algorithm for reducing the effects of multicollinearity on the coefficients is presented not only for linear but for logit and some other linear link functions in the recent work [Lipovetsky (2010b)].

6. Summary

The considered techniques permit to represent conditional, multinomial, and mixed categorical output models as BL regressions. Therefore, it opens a way to use easily available software for logistic regression modeling for constructing more complicated models with multiple choices. Transformation to the binary logistic model permits to reconstruct the coefficients of logit, so that they become free from multicollinearity distortion. It is a useful possibility to produce meaningful coefficients of regression for more complicated conditional and MNL regressions. Consideration of the marginal effects, elasticity, and half-elasticity of these models' probabilities permits to find the adequate interpretation of the coefficients in each model, and even to define the coefficients on the individual observation level that can be used in finding utility contributions from each predictor to their aggregates on the respondent level. The suggested approach is useful for theoretical description and practical applications of discrete choice modeling and analysis.

References

Arminger, G., Clogg, C. C. and Sobel, M. E. (eds.) (1995). *Handbook of Statistical Modeling for the Social and Behavioral Sciences*, Plenum Press, New York.

Ben-Akiva, M. and Lerman, S. R. (1985). *Discrete Choice Analysis*, MIT Press, Cambridge, MA.

Berry, S., Linton, O. and Pakes, A. (2004). Limit theorems for estimating the parameters of differentiated product demand systems. *Rev. Econ. Stud.*, **71**: 613–654.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*, Springer, New York.

Greene, W. H. and Hensher, D. A. (2010). Does scale heterogeneity across individuals matter? An empirical assessment of alternative logit models. *Transportation*, **37**: 413–428.

Gregory, P. (2010). *Bayesian Logical Data Analysis for the Physical Sciences*, Cambridge University Press, Cambridge.

Hastie, T., Tibshirani, R. and Friedman, J. (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer, New York.

Hausman, J. A. and McFadden, D. (1984). A specification test for the multinomial logit model. *Econometrica*, **52**: 1219–1240.

Lipovetsky, S. (2006). Entropy criterion in logistic regression and Shapley value of predictors. *J. Mod. Appl. Stat. Meth.*, **5**: 121–132.

Lipovetsky, S. (2008a). Bradley–Terry choice probability in maximum likelihood and eigenproblem solutions. *Int. J. Inform. Tech. Decis. Making*, **7**: 395–405.

Lipovetsky, S. (2008b). Multinomial structuring in linear regression. *Model Assist. Stat. Appl.*, **3**: 241–247.

Lipovetsky, S. (2009a). Regression with individual coefficients defined via multinomial shares of predictors. *Int. J. Oper. Quant. Manag.*, **15**: 101–116.

Lipovetsky, S. (2009b). Linear regression with special coefficient features attained via parameterization in exponential, logistic, and multinomial-logit forms. *Math. Comput. Model.*, **49**: 1427–1435.

Lipovetsky, S. (2009c). PCA and SVD with nonnegative loadings. *Pattern Recogn.*, **42**: 68–76.

Lipovetsky, S. (2010a). Double logistic curve in regression modeling. *J. Appl. Stat.*, **37**: 1785–1793.

Lipovetsky, S. (2010b). Meaningful regression coefficients built by data gradients. *Adv. Adapt. Data Anal.*, **2**: 451–462.

Lloid, C. J. (1999). *Statistical Analysis of Categorical Data*, Wiley, New York.

Long, J. S. (1997). *Regression Models for Categorical and Limited Dependent Variables*, Sage Publication, London.

Louviere, J. J., Hensher, D. A. and Swait, J. (2000). *Stated Choice Methods: Analysis and Applications*, Cambridge University Press, Cambridge.

McCullagh, P. and Nelder, J. A. (1997). *Generalized Linear Models*, Chapman and Hall, London, New York, pp. 211–212.

McFadden, D. (1973). Conditional logit analysis of qualitative choice behavior, *Frontiers of Econometrics*, Ed. P. Zarembka, Academic Press, New York, pp. 105–142.

McFadden, D. (1981). Econometric models of probabilistic choice. *Structural Analysis of Discrete Data*, Eds. Manski, C. and McFadden, D. MIT Press, Cambridge, pp. 198–272.

McFadden, D. (1984). Econometric analysis of qualitative response models, *Handbook of Econometrics*, Eds. Griliches, Z. and Intriligator, M. Elsevier, North Holland, Amsterdam, Vol. II, pp. 1395–1457.

McFadden, D. and Richter, M. K. (1990). Stochastic rationality and revealed stochastic preference. *Preferences, Uncertainty, and Optimality: Essays in Honor of Leo Huwicz*, Eds. Chipman, J., McFadden D. and Richter, M. K. Westview Press, Boulder, CO, pp. 151–186.

McFadden, D. (2000). *Economic Choices*, Nobel lecture on the microeconometric analysis of choice behavior of consumers who face discrete economic alternatives, http://emlab.berkeley.edu/pub/users/mcfadden/nobel/final-nobel.pdf.

Ripley, B. D. (1997). *Pattern Recognition and Neural Networks*, Cambridge University Press, Cambridge, UK.

Train, K. (2003). *Discrete Choice Methods with Simulation*, Cambridge University Press, New York.

Wedel, M. and Kamakura, W. (1999). *Market Segmentation: Conceptual and Methodological Foundations*, Kluwer Academic Publishers, Dordrecht.