Mathematical and Computer Modelling 49 (2009) 1427–1435

Contents lists available at ScienceDirect

Mathematical and Computer Modelling

journal homepage: www.elsevier.com/locate/mcm

Linear regression with special coefficient features attained via parameterization in exponential, logistic, and multinomial–logit forms

Stan Lipovetsky

GfK Custom Research North America, 8401 Golden Valley Road, Minneapolis, MN 55427, United States

ARTICLE INFO

Article history:
Received 20 February 2008
Received in revised form 3 November 2008
Accepted 12 November 2008

Keywords:
Multiple regression model
Predictors' impact
Exponential
Logistic
Multinomial parameterization

ABSTRACT

Multiple linear regression with special properties of its coefficients parameterized by exponent, logit, and multinomial functions is considered. To obtain always positive coefficients the exponential parameterization is applied. To get coefficients in an assigned range, the logistic parameterization is used. Such coefficients permit us to evaluate the impact of individual predictors in the model. The coefficients obtained by the multinomial–logit parameterization equal the shares of the predictors, which is useful for interpretation of their influence. The considered regression models are constructed by nonlinear optimization techniques, have stable solutions and good quality of fit, have simple structure of the linear aggregates, demonstrate high predictive ability, and suggest a convenient way to identify the main predictors.

1. Introduction

Multiple linear regression is one of the main tools of statistical modeling widely used for estimation of a dependent variable by its predictors. Regressions are very effective for prediction, but are not always useful for the analysis and interpretation of the individual predictors' input due to multicollinearity effects. Multicollinearity distortion of the regression coefficients is well known and described in numerous works, for instance [1–4]. Beginning from a one-parameter ridge-regression approach [5–7], various other techniques have been developed for overcoming the effects of multicollinearity on the coefficients of regression, see, for instance, [8–11]. Among the latest innovations in regression and principal component analyses, a lot of attention has been paid to the regularization methods based on the quadratic L_2-metric, lasso L_1-metric, and other L_p-metrics and their combinations such as elastic net or sparse analysis [12–17].

The current paper considers another approach to constructing a sparse linear combination of the predictors in the regression model using the coefficient parameterization in a special form of exponential, logistic, and multinomial–logit functions. This approach is motivated by necessity to obtain a multiple regression, for instance, with positive coefficients if the pair correlations are positive as well. In many practical problems, particularly, in marketing and advertising research, all the predictors by their meaning should have a definite positive impact on the dependent variable, and it can be easily proven by their pair correlations. However, the coefficients in multiple regression being proportional to the partial correlations could often receive signs opposite to their pair relation signs. Of course, this can be attributed to multicollinearity effects, but it hardly helps in interpretation of the model and in estimation of the individual predictors' contribution.

In these situations, the exponent parameterization of a linear model's coefficients always produces positive coefficients, or coefficients with the signs of their pair correlations. Logistic parameterization can be used to attain all the coefficients in any assigned range of values, for instance from zero to one. Multinomial-logit parameterization yields coefficients with their total equal to one, so such coefficients directly present the shares of the predictors' impact on the response variable.

E-mail address: stan.lipovetsky@gfk.com.

0895-7177/$ – see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.mcm.2008.11.013

Estimation of the parameterized coefficients can be performed by an optimization objective reduced to a Newton–Raphson procedure for nonlinear equations [18–20]. Regressions with special properties of the coefficients can be easier to interpret than ordinary regression models. Such regressions generate stable coefficients of a simple structure in the linear aggregate, demonstrate good prediction ability, and suggest a convenient way to identify the main predictors.

A similar parameterization technique has recently been applied in principal component analysis (PCA) and in singular value decomposition (SVD) to produce loadings with only positive elements, or elements totaling one hundred percent. In contrast to regular PCA and SVD, non-negative loadings have a clear meaning of variable contribution to data approximation and explicitly show which variables with which shares are composed at each step of approximation [21]. Application of the nonlinear parameterization for obtaining only non-negative weights has been considered for sample balance problems in [22].

The paper is arranged as follows. Section 2 presents regressions with several parameterization functions of the coefficients, and describes algorithms for their estimation. Section 3 discusses numerical results, and Section 4 summarizes.

2. Special parameterization of multiple linear regression coefficients

Consider several properties of the ordinary least squares (OLS) regression. A multiple linear regression can be presented as a model:

$$y_i = a_1 x_{i1} + \cdots + a_n x_{in} + \varepsilon_i \equiv \hat{y}_i + \varepsilon_i, \tag{1}$$

where x_{ij} and y_i are centered ith observations ($i = 1, \ldots, N$ — number of observations) by jth independent variables x_j ($j = 1, \ldots, n$ — number of variables) and by the dependent variable y, a_j are the coefficients of regression, $\hat{y}_i$ with hat denotes the theoretical linear aggregate of the predictors, and ε_i are the deviations from the theoretical relationship. The Least Squares (LS) objective minimizes the deviations of the observations from the theoretical model:

$$S^2(a) = \sum_{i=1}^{N} \varepsilon_i^2 = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{N} (y_i - a_1 x_{i1} - \cdots - a_n x_{in})^2 . \tag{2}$$

As is known in regression analysis, the OLS solution can be presented as follows:

$$a_{OLS} = (X'X)^{-1} X'y, \tag{3}$$

where X is the N by n matrix of the centered independent variables, y is the Nth order vector-column of the centered dependent variable, prime denotes transposition, and a_{OLS} is the nth order vector of the estimated regression coefficients (1). With the solution (3) the intercept for the model in the original variables is found: $a_0 = \bar{y} - a_1 \bar{x}_1 - \cdots - a_n \bar{x}_n$. If the data are centered and normalized by the standard deviations, then solution (3) yields the so-called beta-coefficients β of the standardized model, and then the coefficients of the original model are defined as $a_j = \beta_j \sigma_y / \sigma_j$, where σ_j and σ_y are the standard deviations of the variables x_j and y, respectively. In the case of multicollinearity, some coefficients of the regression receive signs opposite to the signs of their correspondent pair correlations with the dependent variable.

In numerous applied regression problems – for instance, in marketing research – the direction of pair relations among the variables can be known a priori by their meaning. Suppose, all the pair relations should be positive, and it is verified by the pair correlations. If not, it is always possible to invert the variable scale to obtain all positive pair relations. The problem is — how to obtain a multiple regression model where each regressor exhibits a positive influence on the dependent variable? An easy and convenient way to obtain such a model is suggested by the nonlinear parameterization of the regression coefficients.

If all non-negative coefficients are sought, they can be presented in the exponential parameterization:

$$a_j = \exp(\gamma_j), \tag{4}$$

where γ_j are the estimated parameters. To obtain the coefficients of regression belonging to any given span of values from $a_{\min}$ to $a_{\max}$, a logistic parameterization can be applied:

$$a_j = a_{\min} + \frac{a_{\max} - a_{\min}}{1 + \exp(-\gamma_j)}. \tag{5}$$

For instance, with the constants $a_{\min} = 0$ and $a_{\max} = 1$ each coefficient of regression would belong to the [0, 1] interval. The multinomial–logit parameterization

$$a_j = \frac{\exp(\gamma_j)}{\exp(\gamma_1) + \exp(\gamma_2) + \cdots + \exp(\gamma_n)}, \quad \gamma_1 = 0, \tag{6}$$

produces all non-negative coefficients of regression with their total equal to one. One of the parameters in (6) is redundant and can be put to zero, for instance, $\gamma_1 = 0$. In contrast to the exponent or logistic parameterization (4) and (5) where each coefficient of regression a_j depends on just one corresponding parameter γ_j, in the multinomial parameterization each coefficient of regression a_j is a function of all $n - 1$ free parameters γ_j. If all the variables in the regression are measured in the same scale, it is possible to apply any parameterization of (4)–(6) directly to the coefficients of regression a_j. If the