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

Table 1
Marketing data: pair correlations.

	y	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}	x_{14}
y	1.00	0.50	0.49	0.40	0.48	0.35	0.45	0.57	0.50	0.49	0.41	0.32	0.29	0.45	0.32
x_1	0.50	1.00	0.61	0.53	0.52	0.52	0.57	0.62	0.58	0.52	0.58	0.49	0.41	0.45	0.49
x_2	0.49	0.61	1.00	0.47	0.69	0.38	0.63	0.70	0.51	0.57	0.52	0.47	0.42	0.43	0.48
x_3	0.40	0.53	0.47	1.00	0.48	0.55	0.51	0.58	0.59	0.58	0.63	0.48	0.57	0.48	0.53
x_4	0.48	0.52	0.69	0.48	1.00	0.35	0.67	0.70	0.55	0.61	0.48	0.40	0.58	0.38	0.58
x_5	0.35	0.52	0.38	0.55	0.35	1.00	0.37	0.48	0.51	0.39	0.71	0.60	0.35	0.51	0.39
x_6	0.45	0.57	0.63	0.51	0.67	0.37	1.00	0.70	0.55	0.71	0.47	0.40	0.53	0.41	0.53
x_7	0.57	0.62	0.70	0.58	0.70	0.48	0.70	1.00	0.58	0.65	0.62	0.50	0.50	0.48	0.54
x_8	0.50	0.58	0.51	0.59	0.55	0.51	0.55	0.58	1.00	0.56	0.57	0.48	0.55	0.48	0.52
x_9	0.49	0.52	0.57	0.58	0.61	0.39	0.71	0.65	0.56	1.00	0.50	0.40	0.58	0.40	0.52
x_{10}	0.41	0.58	0.52	0.63	0.48	0.71	0.47	0.62	0.57	0.50	1.00	0.65	0.42	0.48	0.47
x_{11}	0.32	0.49	0.47	0.48	0.40	0.60	0.40	0.50	0.48	0.40	0.65	1.00	0.39	0.43	0.45
x_{12}	0.29	0.41	0.42	0.57	0.58	0.35	0.53	0.50	0.55	0.58	0.42	0.39	1.00	0.36	0.63
x_{13}	0.45	0.45	0.43	0.48	0.38	0.51	0.41	0.48	0.48	0.40	0.48	0.43	0.36	1.00	0.42
x_{14}	0.32	0.49	0.48	0.53	0.58	0.39	0.53	0.54	0.52	0.52	0.47	0.45	0.63	0.42	1.00

scales are different, each parameterization should be applied to the β_j coefficients of the standardized regression, and then the original regression coefficients are found similarly to the OLS model.

When the vector of regression coefficients is a function $a = a(\gamma)$ of the vector of free parameters γ, the LS objective (2) becomes:

$$S^2(\gamma) = \sum_{i=1}^{N} \varepsilon_i^2 = \sum_{i=1}^{N} (y_i - a_1(\gamma)x_{i1} - \cdots - a_n(\gamma)x_{in})^2 . \tag{7}$$

Numerical minimization of a nonlinear objective (7) can be performed by utilizing techniques available in modern statistical packages. In explicit form such a minimization for the parameterizations (4)–(6) can be presented by the Newton–Raphson algorithm — see Appendix. The iterative processes for minimizing the objective (7) with the parameterization (4), (5) and (6) are presented explicitly in the formulae (A.8), (A.10) and (A.20), respectively.

The parameterization of the regression coefficients can be applied through various modifications. If a positive intercept is required, this can be parameterized too. If all the variables are measured in the same scale, and if the minimum and maximum of a predicted value should belong to a specific range (for instance, of the same scale) then the intercept should be used as one of the shared coefficients in the multinomial parameterization. The signs of the pair correlations of the x values with y can be incorporated into the estimated parameters. If the variables are measured in different units, the modeling should be performed by the standardized data with subsequent transition to the original units. Statistical significance for the parameterized coefficients of regression can be found via the diagonal elements of the inverse Hessian matrix, which present estimates of the coefficients' variance, or, even more easily, using bootstrap resampling evaluation. The t-statistic for a regression parameter is a quotient of the coefficient and its standard deviation, $t_j = a_j/\sigma(a_j)$. More complicated quality characteristics for nonlinear models and measures of difference between OLS and non-OLS models are considered in [8,10,11,18–20,26–28].

3. Numerical results

The first example is taken from a real marketing research project on a food product, with 720 respondents using a seven-point scale from "do not agree at all" to "fully agree". The evaluated variables are: y — uniqueness, x_1 — good value for money, x_2 — taste, x_3 — innovative, x_4 — product usage, x_5 — healthy lifestyle, x_6 — premium, x_7 — feel good about it, x_8 — modern, x_9 — leader among the brands, x_{10} — wholesome, x_{11} — natural, x_{12} — wide variety, x_{13} — advertising, x_{14} — convenient meal. The dependent variable was modeled in a regression by all the other variables to find the key drivers of the product uniqueness.

Table 1 presents the pair correlations of y and x values, all of which are positive and spread from a minimum at 0.29 to a maximum at 0.71, with a mean value of 0.47. The variance inflation factor reaches a maximum of 3.28, so the predictors are only mildly collinear. However, this is sufficient to produce change in the predictors' signs in multiple regression. Table 2 contains the model results by several methods — ordinary least squares (OLS), the stepwise reduced variant of the OLS model, Shapley Value (SV) regression described in [10], common ridge regression with one parameter (ridge-1) [5,6], and its modification in ridge regression with two parameters (ridge-2) [27].

In the first two numerical columns, Table 2 shows the coefficients and their t-statistics for OLS regression. All the predictors by their meaning are expected to have positive relations with the dependent variable, which is supported by their positive pair correlations. However, seven of the 14 predictors have negative signs of their coefficients in the multiple linear regression. The total of all the predictors' coefficients (without the intercept related to the identity variable x_0) and the coefficient of multiple determination ($R^2 = 0.44$) are shown in the two bottom rows. Several t-values of the regression coefficients are close to zero, and a stepwise elimination of the insignificant variables (forward–backward procedure based on partial F-statistics) yields a reduced model shown in the next two columns of Table 2. The stepwise model keeps nine

Table 2
Marketing data: OLS, stepwise, Shapley-Value, ridge-1, and ridge-2 regressions.

	OLS model		Stepwise model		SV regression		Ridge regressions		
							Ridge-1	Ridge-2	
Item	a_j	t_j	a_j	t_j	a_j	t_j	a_j	a_j	t_j
x_0	0.83	3.68	0.80	3.63	0.39	1.77	1.46	0.35	
x_1	0.16	3.59	0.16	3.74	0.09	4.34	0.07	0.10	4.35
x_2	0.05	1.08			0.07	4.72	0.06	0.08	4.04
x_3	−0.01	−0.15			0.04	3.45	0.03	0.04	2.08
x_4	0.13	2.45	0.14	2.87	0.08	4.74	0.06	0.08	3.83
x_5	−0.02	−0.39			0.02	2.79	0.02	0.03	1.70
x_6	−0.08	−1.38			0.06	5.58	0.05	0.06	2.94
x_7	0.26	4.72	0.25	4.88	0.12	6.30	0.08	0.11	5.76
x_8	0.21	4.69	0.20	4.61	0.10	4.62	0.07	0.10	4.46
x_9	0.21	3.97	0.18	0.85	0.09	4.28	0.07	0.09	4.21
x_{10}	−0.02	−0.34			0.03	3.91	0.03	0.04	2.28
x_{11}	−0.05	−1.25	−0.05	−1.59	0.01	1.06	0.01	0.02	0.90
x_{12}	−0.15	−3.04	−0.16	−3.34	0.02	1.03	0.00	0.00	0.05
x_{13}	0.18	5.37	0.17	5.45	0.09	4.35	0.06	0.08	4.11
x_{14}	−0.10	−2.37	−0.11	−2.47	0.02	1.42	0.01	0.01	0.47
Total	0.78		0.79		0.85		0.64	0.86	
R^2	0.44		0.44		0.39		0.36	0.39	

Table 3
Marketing data: regressions with exponential, logistic, and multinomial–logit parameterization.

	Exponential parameterization		Logistic parameterization		Multinomial parameterization	
Item	a_j	t_j	a_j	t_j	a_j	t_j
x_0	0.40	1.59	0.39	1.61	−0.32	−2.68
x_1	0.16	2.01	0.13	2.05	0.16	2.28
x_2	0.11	1.15	0.06	1.24	0.07	1.49
x_3						
x_4	0.10	0.89	0.04	0.96	0.07	1.28
x_5						
x_6						
x_7	0.01	1.59	0.22	2.97	0.22	2.76
x_8	0.16	2.30	0.14	2.32	0.16	2.67
x_9	0.17	1.73	0.12	1.68	0.16	2.08
x_{10}						
x_{11}						
x_{12}						
x_{13}	0.16	3.69	0.14	3.67	0.16	5.26
x_{14}						
Total	0.87		0.86		1.00	
R^2	0.40		0.41		0.40	

significant variables with three of them still of negative sign for the coefficients. The coefficient of multiple determination for the reduced model shows practically the same quality of fit as in the original model. The negative coefficients in these models can be interpreted as the effects of multicollinearity, although this doesn't help much in estimation of the individual predictors' impact on the dependent variable.

In the next two columns of Table 2, the coefficients and their t-statistics for SV regression are presented. The results of SV consist of the coefficients in multiple regression but with signs of the pair relations, so the shares of the individual predictors are easily interpretable. The t-statistics of the coefficients are mostly high, so this model is very robust. The coefficient of multiple determination in the SV model is $R^2 = 0.39$, which is slightly lower than in the OLS model — this is the price of adjusting to the interpretable coefficients. The last three columns of Table 2 contain the results of ridge-1 and ridge-2 regressions (taken at the level of the profile parameter $k = 2.2$ when all the coefficients became positive). These solutions are proportional (with the second parameter $q = 1.35$), so they have the same t-statistics, but the coefficient of multiple determination in ridge-2 is always higher than in ridge-1 ($R^2 = 0.39$ and $R^2 = 0.36$, respectively). The SV and ridge-2 results are rather similar.

Table 3 presents modeling by the suggested special parameterizations of the regression coefficients, including the coefficients and their t-statistics for the exponential, logistic, and multinomial–logit parameterizations (4)–(6). The logit parameterization (5) is considered for the interval [0, 1], and the model is slightly better than the exponential parameterization model ($R^2 = 0.41$ and $R^2 = 0.40$, respectively). The last two columns in Table 3 present the results of the multinomial parameterization (6) for the coefficients of the linear model (1). The structure of the coefficients of the multinomial parameterization is similar to those of the exponential or logistic models (note that absence of some coefficients means their zero values). However, the multinomial approach yields coefficients (without the intercept) that sum to one. It

Table 4
Marketing data: bootstrapping for models and predictions.

	OLS regression				Logit parameterization				Multinomial parameterization			
Item	min a_j	max a_j	mean a_j	t_j	min a_j	max a_j	mean a_j	t_j	min a_j	max a_j	mean a_j	t_j
x_0	.13	1.4	.75	3.02	−.37	.80	.34	1.48	−1.06	−.25	−.55	−3.21
x_1	.04	.29	.15	2.56	0	.24	.12	2.03	0	.26	.14	2.33
x_2	−.08	.26	.06	1.00	0	.25	.07	1.26	0	.28	.10	1.51
x_3	−.10	.12	.00	.00	0	.05	.00	.20	0	.08	.00	.32
x_4	−.04	.25	.12	2.06	0	.18	.04	.99	0	.22	.06	1.22
x_5	−.12	.09	−.01	−.25	0	.06	.00	.18	0	.08	.01	.62
x_6	−.24	.08	−.06	−.99	0	.05	.00	.29	0	.07	.01	.35
x_7	.05	.45	.25	3.36	.04	.39	.22	2.88	0	.39	.21	2.77
x_8	.06	.35	.22	3.70	0	.25	.15	2.72	.03	.28	.17	3.15
x_9	.06	.36	.20	3.13	0	.25	.12	2.01	0	.29	.15	2.56
x_{10}	−.13	.11	−.01	−.26	0	.05	.00	.19	0	.07	.00	.32
x_{11}	−.13	.03	−.05	−1.17	0	0	0	0	0	.04	.00	.18
x_{12}	−.29	.00	−.14	−2.24	0	0	0	0	0	0	0	0
x_{13}	.07	.29	.17	4.45	0	.25	.14	3.45	.07	.29	.18	4.68
x_{14}	−.20	.02	−.10	−2.07	0	0	0	0	0	0	0	0
STD sample	1.13	1.34	1.26		1.17	1.39	1.29		1.18	1.42	1.31	
R^2 sample	.38	.54	.46		.35	.52	.43		.32	.51	.41	
STD predict	1.26	1.45	1.34		1.25	1.48	1.35		1.27	1.49	1.38	
R^2 predict	.27	.47	.39		.27	.45	.38		.23	.45	.35	

is the only model (see the row above the last one in Tables 2 and 3) with coefficients coinciding with the shares of the impact of each individual predictor on the dependent variable. The coefficient of multiple determination in this model $R^2 = 0.40$ is similar to the values of R^2 in the other models with parameterized coefficients, a little lower than in the OLS and stepwise models, but higher than in the SV and ridge regressions (see Table 2). So the new approaches (in Table 3) yield interpretable coefficients and better fit with the data than the SV and ridge regressions (in Table 2). The sparse structure of the key drivers due to all three parameterizations in Table 3 is very convenient for identification of the most useful variables in the model — those are x_1, x_2, x_4, x_7, x_8, x_9, and x_{13}.

Next, simulation via re-sampling is applied to estimate the stability of the regression results and their predictive ability. In 150 random samplings, about a half of the observations were picked up and used to construct regressions by several considered techniques. Then the obtained models were applied to predict the output y by the x values for the other half of the data. Table 4 presents the bootstrapping results for modeling and forecasting by three main regressions. The OLS, logistic and multinomial–logit parameterized models were constructed 150 times by each of the random samples — the descriptive statistics of the results are presented in Table 4. For each coefficient of regression, Table 4 shows its minimum (*min*), maximum (*max*), mean value, and also the t-statistics for a coefficient's mean. Below the coefficients of regression, Table 4 presents the descriptive statistics for the standard residual variance and the coefficient of multiple determination of the models constructed by the random samples (*STD* sample, and R^2 sample). The last two rows in Table 4 show the descriptive statistics for the standard residual variance and the coefficient of multiple determination for the predicted values of the dependent variable compared with the empirical values (*STD* predicted, and R^2 predicted).

The mean values of R^2 in the sampled data are 0.46, 0.43, and 0.41, and in the predicted data they are 0.39, 0.38, and 0.35, for the OLS, logit, and multinomial parameterization, respectively. The results across the models are similar, although slightly diminishing from the first to the last of the models. It is quite expected, because the more restrictive models get a lower quality of fit and predictive power. But in contrast to the OLS model, the logit and multinomial parameterization yield regressions with all interpretable coefficients. It is interesting to note that the minimum values of the coefficients by the OLS models (the first numerical column in Table 4) are positive only for the predictors x_1, x_7, x_8, x_9, and x_{13} — the same variables indicated above among the main key drivers. Bootstrapping results in Table 4 support this choice of the key drivers, and show that the large positive t-statistics are reached exactly by these variables in all the estimations.

The second example is considered by the car data from [29], also available in [30] ("cu.summary" file). The data contain dimensions and mechanical specifications of 111 various cars, supplied by the manufacturers and measured by Consumer Reports. The variables taken in the example are: y — price of a car, US$K; x_1 — weight, pounds; x_2 — length overall, inches; x_3 — wheel base length, inches; x_4 — width, inches; x_5 — front leg room maximum, inches; x_6 — front shoulder room, inches; x_7 — turning circle radius, feet; x_8 — displacement of the engine, cubic inches; x_9 — HP, the net horsepower; x_{10} — tank fuel refill capacity, gallons. The cars' price is estimated in the regression model by the dimensions and specification variables.

Table 5 presents pair correlations of the variables in the car data. All the values are positive, and correlations among the predictors are often higher than their correlations with the dependent variable. It is a situation prone to multicollinearity effects [4]. Indeed, Table 6 in the first two numerical columns presents the OLS coefficients. Four out of ten predictors have a negative relationship with the dependent variable. Besides the coefficients a_j in the original units, the standardized beta-coefficients b_j of regression are also given, together with their total at the bottom, and the coefficients of multiple

Table 5
Car data: pair correlations.

	Price	Weight	Length	Wheel base	Width	Frt. Leg Room	Frt. Shld	Turn.	Disp.	HP	Tank
Price	1.00	0.65	0.53	0.50	0.48	0.57	0.37	0.38	0.64	0.78	0.66
Weight	0.65	1.00	0.81	0.80	0.86	0.31	0.82	0.77	0.85	0.70	0.85
Length	0.53	0.81	1.00	0.82	0.79	0.25	0.75	0.79	0.78	0.53	0.68
Wheel base	0.50	0.80	0.82	1.00	0.79	0.34	0.76	0.78	0.73	0.52	0.75
Width	0.48	0.86	0.79	0.79	1.00	0.26	0.89	0.78	0.83	0.52	0.78
Frt. Leg. Room	0.57	0.31	0.25	0.34	0.26	1.00	0.12	0.14	0.31	0.45	0.40
Frt. Shld	0.37	0.82	0.75	0.76	0.89	0.12	1.00	0.71	0.72	0.38	0.71
Turning	0.38	0.77	0.79	0.78	0.78	0.14	0.71	1.00	0.75	0.47	0.65
Disp.	0.64	0.85	0.78	0.73	0.83	0.31	0.72	0.75	1.00	0.75	0.77
HP	0.78	0.70	0.53	0.52	0.52	0.45	0.38	0.47	0.75	1.00	0.68
Tank	0.66	0.85	0.68	0.75	0.78	0.40	0.71	0.65	0.77	0.68	1.00

Table 6
Car data: regression models by several types of estimation.

	OLS regression		SV regression		Ridge-2 regression		Exponential parameters		Multinomial–logit parameters	
	a_j	b_j	a_j	b_j	a_j	b_j	a_j	b_j	a_j	b_j
Intercept	−85.18		−90.38		−109.21		−129.03		−125.92	
Weight	0.00	0.28	0.00	0.13	0.00	0.13	0.00	0.09		
Length	0.13	0.22	0.04	0.07	0.05	0.08	0.04	0.07	0.06	0.10
Wheel. base	−0.10	−0.08	0.05	0.04	0.06	0.05				
Width	−0.36	−0.14	0.12	0.05	0.09	0.04				
Frt. Lg. Rm	2.69	0.25	1.54	0.14	2.01	0.18	2.83	0.26	2.72	0.25
Frt. Shld	−0.18	−0.06	0.04	0.01	0.01	0.00				
Turning	−0.52	−0.20	0.06	0.02	0.01	0.00				
Disp.	0.01	0.10	0.02	0.11	0.02	0.12				
HP	0.08	0.41	0.04	0.19	0.05	0.24	0.10	0.51	0.11	0.53
Tank	0.41	0.16	0.29	0.11	0.35	0.14	0.22	0.08	0.32	0.13
Total		0.93		0.88		0.98		1.01		1.00
R2		0.72		0.60		0.63		0.70		0.69

determination are shown in the last row of Table 6. Similarly to Table 2, after the OLS results Table 6 presents two columns of the coefficients a_j in the original units and in the standardized beta-coefficients b_j for the SV and ridge-2 regressions. Then similarly to Table 3, the results of exponential and multinomial–logit parameterization are shown in Table 6. In the car data example the variables have different units, so the parameterized modeling is performed for the standardized variables. The beta-coefficients b_j are estimated first, then they are transformed to the coefficients in the original units, as discussed after formula (3).

Comparison of the models in Table 6 demonstrates that all the SV, ridge-2, and parameterized techniques produce only non-negative coefficients, with clear, sparse structure in the two latter approaches. The total of the beta-coefficients equals exactly one only for the multinomial–logit model, which suggests the coefficients equivalent to the shares of the predictors' impact on the dependent variable. This model also identifies the main price leading predictors as follows: length, front leg room, HP, and tank capacity. The coefficients of multiple determination given in the last row of Table 6 show that the exponential and multinomial–logit parameterizations outperform the other techniques designed for obtaining interpretable regression models.

The described numerical results are very typical and have been observed by numerous models obtained with different data sets for various research applications. Indeed, the SV and ridge regressions try to construct all the coefficients with interpretable signs, but lose more on the side of quality of fit than the other models with the parameterized coefficients which lead to zero values of the least important variables and reveal the simple sparse structure of the regression key drivers.

4. Summary

Models linear by their variables but nonlinear by the estimated coefficients are considered. To always get positive coefficients, an exponential parameterization is used, and for coefficients belonging to a specified range, a logistic parameterization is applied. The coefficients equal to the shares of the predictors' impact on the dependent variable are considered in multinomial parameterization. Nonlinear optimizing procedures are implemented for the estimations. The suggested models demonstrate stable, adequate, and more interpretable results than the results of the ordinary least squares regression. They also have a good quality of fit and high predictive power. A simple structure in the linear aggregate with zero-coefficients for the unimportant variables is obtained. The proposed techniques are convenient for finding the key drivers, and are helpful in practical applications of statistical modeling.

Acknowledgements

The author thanks two reviewers for their valuable comments which helped to improve the paper.

Appendix

The Newton–Raphson algorithm considers the objective (7) approximation as:

$$S^2(\gamma) \approx S^2(\gamma^{(0)}) + \frac{\partial S^2}{\partial \gamma}(\gamma - \gamma^{(0)}), \tag{A.1}$$

where $\gamma^{(0)}$ is an initial approximation to the vector of the parameters. Minimization of the function corresponds to the condition of the first derivative equal to zero, so taking the derivative of (A.1) yields:

$$\frac{\mathrm{d}S^2}{\mathrm{d}\gamma} = \frac{\partial^2 S^2}{\partial \gamma \partial \gamma'}(\gamma - \gamma^{(0)}) + \frac{\partial S^2}{\partial \gamma} = 0. \tag{A.2}$$

The solution of equation (A.2) for the vector γ is:

$$\gamma = \gamma^{(0)} - \left(\frac{\partial^2 S^2}{\partial \gamma \partial \gamma'}\right)^{-1} \left(\frac{\partial S^2}{\partial \gamma}\right). \tag{A.3}$$

The expression (A.3) is used in iterations for each $(t+1)$th approximation of vector $\gamma^{(t+1)}$ via the previous vector $\gamma^{(t)}$ at the tth step, which is the Newton–Raphson procedure for nonlinear equations. Consider applications of this procedure to the parameterizations (4)–(6).

In the parameterization (4) and (5), the first derivatives of (7) with respect to the parameters are:

$$U_k = \frac{\partial S^2}{\partial \gamma_k} = -2\sum_{i=1}^{N} x_{ik}\left(y_i - \sum_{j=1}^{n} a_j x_{ij}\right)\left(\frac{\mathrm{d}a_k}{\mathrm{d}\gamma_k}\right). \tag{A.4}$$

The second derivatives are:

$$H_{qk} = \frac{\partial U_k}{\partial \gamma_q} = \frac{\partial^2 S^2}{\partial \gamma_q \partial \gamma_k} = 2\frac{\mathrm{d}a_q}{\mathrm{d}\gamma_q}\frac{\mathrm{d}a_k}{\mathrm{d}\gamma_k}\sum_{i=1}^{N} x_{iq}x_{ik} - 2\delta_{qk}\frac{\mathrm{d}^2 a_k}{\mathrm{d}\gamma_k^2}\sum_{i=1}^{N} x_{ik}\left(y_i - \sum_{j=1}^{n} a_j x_{ij}\right), \tag{A.5}$$

where δ_{qk} is the Kronecker delta. The second sum in (A.5) contains the predictors x_{ik} multiplied by residual errors ε_i, or scalar products of the vectors of errors and predictors. Such vectors are approximately orthogonal, so their scalar products are close to zero, and the last item with the term δ_{qk} in (A.5) can be omitted. Such a simplification of a Hessian, or a matrix of second derivatives (A.5), makes it more robust for inversion in (A.3) for the iterative process, and is recommended for numerical estimations [20–24].

In matrix notations, the gradient vector with the elements (A.4) and the simplified Hessian (A.5) can be represented as:

$$U = -2DX'\varepsilon, \qquad H = 2D(X'X)D, \qquad D = \mathrm{diag}\left(\frac{\mathrm{d}a}{\mathrm{d}\gamma}\right), \tag{A.6}$$

where X is a matrix of predictors, ε is the vector of residuals (1), and D is a diagonal matrix of derivatives of the regression coefficients by parameters γ. Using (A.6) in (A.3) yields:

$$\begin{aligned}\gamma^{(t+1)} = \gamma^{(t)} + D^{-1}(X'X)^{-1}D^{-1}DX'(y - Xa) &= \gamma^{(t)} + D^{-1}\left((X'X)^{-1}X'y - a\right) \\ &= \gamma^{(t)} + D^{-1}(a_{OLS} - a^{(t)}),\end{aligned} \tag{A.7}$$

where the OLS solution is defined in (3), and $a^{(t)}$ is the iterated vector of coefficients (4) or (5). The Newton–Raphson iteration process (A.7) usually quickly converges, and with the obtained parameters γ the coefficients of regression a are found by (4) or (5).

A derivative D_k (A.6) for the exponents (4) is $\exp(\gamma_k) = a_k$, so (A.7) reduces to:

$$\gamma^{(t+1)} = \gamma^{(t)} + \exp(-\gamma^{(t)})a_{OLS} - e_n, \tag{A.8}$$

where e_n is the nth order identity vector-column. Similarly for logistic (5), for instance, in the interval [0,1], the derivatives (A.6) are:

$$D_k = \exp(-\gamma_k)/\left(1 + \exp(-\gamma_k)\right)^2 = a_k(1 - a_k), \tag{A.9}$$

and the process (A.7) can be presented explicitly as:

$$\gamma^{(t+1)} = \gamma^{(t)} + \left(1 + \exp(\gamma^{(t)})\right)\left\{\left(1 + \exp(-\gamma^{(t)})\right)a_{OLS} - e_n\right\}. \tag{A.10}$$

For the multinomial parameterization (6), the derivation of the Newton–Raphson procedure is slightly different because each coefficient depends on $n-1$ free parameters $\gamma_2, \ldots, \gamma_n$. Substituting the expression $a_1 = 1 - a_2 - \cdots - a_n$ for the first coefficient via all the others into the objective (7) reduces it to:

$$S^2(\gamma) = \sum_{i=1}^{N} \varepsilon_i^2 = \sum_{i=1}^{N} (\breve{y}_i - a_2(\gamma)\,\breve{x}_{i2} - \cdots - a_n(\gamma)\,\breve{x}_{in})^2\,, \tag{A.11}$$

where the new standard variables are defined as the previous ones minus the first variable:

$$\breve{y}_i = y_i - x_{i1}, \qquad \breve{x}_{i2} = x_{i2} - x_{i1}, \ldots, \breve{x}_{in} = x_{in} - x_{i1}. \tag{A.12}$$

A derivative of a regression coefficient (6) by any parameter is:

$$\frac{\partial a_j}{\partial \gamma_k} = \frac{\delta_{jk}\exp(\gamma_j)}{1+\exp(\gamma_2)+\cdots+\exp(\gamma_n)} - \frac{\exp(\gamma_j)\exp(\gamma_k)}{(1+\exp(\gamma_2)+\cdots+\exp(\gamma_n))^2} = a_k(\delta_{jk} - a_j), \tag{A.13}$$

where δ_{jk} is the Kronecker delta. The first derivatives of (A.11) are now:

$$\begin{aligned} U_k = \frac{\partial S^2}{\partial \gamma_k} &= -2\sum_{i=1}^{N}\left(\breve{y}_i - \sum_{j=1}^{n} a_j\,\breve{x}_{ij}\right)\left(\sum_{j=1}^{n}\breve{x}_{ij}\frac{\partial a_j}{\partial \gamma_k}\right) \\ &= -2\sum_{i=1}^{N}\left(\breve{y}_i - \sum_{j=1}^{n} a_j\,\breve{x}_{ij}\right)\left(a_k\sum_{j=1}^{n}\breve{x}_{ij}(\delta_{jk} - a_j)\right) \\ &= -2a_k\sum_{i=1}^{N}\left(\breve{y}_i - \hat{y}_i\right)\left(\breve{x}_{ik} - \hat{y}_i\right), \end{aligned} \tag{A.14}$$

where the theoretical values $\hat{y}_i = a_2(\gamma)\,\breve{x}_{i2} + \cdots + a_n(\gamma)\,\breve{x}_{in}$ are defined as in (A.11). The second derivatives can be presented as follows:

$$\begin{aligned} H_{qk} = \frac{\partial U_k}{\partial \gamma_q} = \frac{\partial^2 S^2}{\partial \gamma_q \partial \gamma_k} &= 2a_q a_k \sum_{i=1}^{N}(\breve{x}_{iq} - \hat{y}_i)(\breve{x}_{ik} - \hat{y}_i) \\ &-2a_q(\delta_{qk} - a_k)\sum_{j=1}^{n}(\delta_{jk} - a_j)\sum_{i=1}^{N}(\breve{y}_i - \hat{y}_i)\,\breve{x}_{ij} + 2a_q a_k \sum_{j=1}^{n}(\delta_{jq} - a_j)\sum_{i=1}^{N}(\breve{y}_i - \hat{y}_i)\,\breve{x}_{ij}. \end{aligned} \tag{A.15}$$

In (A.14) and (A.15) the indices k and q run from 2 to n. Similarly to (A.5), the last two sums (A.15) contain the scalar products of the error by predictor vectors, so they are close to zero and can be omitted to make the Hessian robust.

In matrix notations the gradient vector (A.14) becomes:

$$\begin{aligned} U &= -2\,\mathrm{diag}(a)(\breve{X} - \hat{y}e'_{n-1})'\varepsilon \\ &= -2\,\mathrm{diag}(a)\left(\breve{X}(I - ae'_{n-1})\right)'\varepsilon \\ &= -2\,\mathrm{diag}(a)(I - e_n a')\breve{X}'\varepsilon \end{aligned} \tag{A.16}$$

where $\mathrm{diag}(a)$ is a diagonal matrix of $n-1$ order of the regression coefficients (6) without the first one, I denotes the $n-1$ order identity matrix, $\hat{y}e'_{n-1}$ is an N by $n-1$ matrix of the outer product of the theoretical vector-column $\hat{y} = \breve{X}a$ by the identity vector-row e'_{n-1} of $n-1$ order, $\breve{X}$ is the N by $n-1$ matrix of the shifted predictors (A.12), and $\varepsilon = \breve{y} - \hat{y}$ is the vector of residuals in (A.11). The simplified Hessian, which is the first sum in (A.15), becomes:

$$\begin{aligned} H &= 2\,\mathrm{diag}(a)(\breve{X} - \hat{y}e'_{n-1})'(\breve{X} - \hat{y}e'_{n-1})\mathrm{diag}(a) \\ &= 2\,\mathrm{diag}(a)\left(I - e_{n-1}a'\right)(\breve{X}'\breve{X})(I - ae'_{n-1})\mathrm{diag}(a). \end{aligned} \tag{A.17}$$

Then the expressions (A.16) and (A.17) can be represented as:

$$U = -2D\breve{X}'\varepsilon, \qquad H = 2D(\breve{X}'\breve{X})D, \qquad D = \mathrm{diag}(a) - aa', \tag{A.18}$$

where D is the $n-1$ order diagonal matrix of the regression coefficients $a_2, \ldots, a_n$ (6) minus the outer product of the vector of these coefficients. Applying the known formula [25] for inversion of a matrix of a structure D (A.18), yields:

$$D^{-1} = \left(\mathrm{diag}(a) - aa'\right)^{-1} = \mathrm{diag}(a^{-1}) + \frac{\mathrm{diag}(a^{-1})aa'\mathrm{diag}(a^{-1})}{1 - a'\mathrm{diag}(a^{-1})a} = \mathrm{diag}(a^{-1}) + \frac{e_{n-1}e'_{n-1}}{1 - a'e_{n-1}}, \tag{A.19}$$

where e_{n-1} and a are the $n-1$ order identity vector and vector of coefficients. Substituting (A.18) and (A.19) into the Newton–Raphson formula (A.3) yields the expression (A.7) which reduces to:

$$\gamma^{(t+1)} = \gamma^{(t)} + D^{-1}(\breve{a}_{OLS} - a^{(t)}) = \gamma^{(t)} + D^{-1}\breve{a}_{OLS} - \frac{1}{1 - e'_{n-1}a^{(t)}}e_{n-1}, \tag{A.20}$$

where $\breve{a}_{OLS} = (\breve{X}'\breve{X})^{-1}\breve{X}'\breve{y}$ is the OLS solution (3) for the variables (A.12), and $a^{(t)}$ is the iterated vector of the expressions (6) without the first coefficient. There are only $n-1$ free parameters γ_j in the multinomial parameterization, with the first one taken as $\gamma_1 = 0$, and at each iteration step the first coefficient of regression (6) can be found by the relation $a_1^{(t)} = 1 - a_2^{(t)} - \cdots - a_n^{(t)}$. So the expression in the denominator (A.20) equals the first coefficient $a_1^{(t)} = 1 - e'_{n-1}a^{(t)}$ at each step of the iterations.

References

[1] A. Grapentine, Managing multicollinearity, Marketing Research 9 (1997) 11–21.
[2] C.H. Mason, W.D. Perreault, Collinearity, power, and interpretation of multiple regression analysis, Journal of Marketing Research 28 (1991) 268–280.
[3] S. Weisberg, Applied Linear Regression, Wiley, New York, 1985.
[4] S. Lipovetsky, M. Conklin, A model for considering multicollinearity, International Journal of Mathematical Education in Science and Technology 34 (2003) 771–777.
[5] A.E. Hoerl, R.W. Kennard, Ridge regression: Biased estimation for nonorthogonal problems, Technometrics 12 (1970) 55–67.
[6] A.E. Hoerl, R.W. Kennard, K.F. Baldwin, Ridge regression: Some simulation, Communications in Statistics A4 (1975) 105–124.
[7] P.J. Brown, Measurement, Regression, Calibration, Oxford University Press, Oxford, 1994.
[8] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning: Data Mining, Inference and Prediction, Springer, New York, 2001.
[9] S. Lipovetsky, M. Conklin, Multiobjective regression modifications for collinearity, Computers and Operations Research 28 (2001) 1333–1345.
[10] S. Lipovetsky, M. Conklin, Analysis of regression in game theory approach, Applied Stochastic Models in Business and Industry 17 (2001) 319–330.
[11] S. Lipovetsky, M. Conklin, Dual- and triple-mode matrix approximation and regression modelling, Applied Stochastic Models in Business and Industry 19 (2003) 291–301.
[12] R. Tibshirani, Regression shrinkage and selection via the lasso, Journal of the Royal Statistical Society B58 (1996) 267–288.
[13] I.T. Jolliffe, N. Trendafilov, M. Uddin, A modified principal component technique based on the lasso, Journal of Computational and Graphical Statistics 12 (2003) 531–547.
[14] B. Efron, T. Hastie, I. Johnstone, R. Tibshirani, Least angle regression, The Annals of Statistics 32 (2004) 407–489.
[15] H. Zou, T. Hastie, R. Tibshirani, Sparse principal component analysis, Journal of Computational and Graphical Statistics 15 (2006) 265–286.
[16] S. Lipovetsky, Optimal Lp-metric for minimizing powered deviations in regression, Journal of Modern Applied Statistical Methods 6 (2007) 219–227.
[17] S. Lipovetsky, Equidistant regression modeling, Model Assisted Statistics and Applications 2 (2007) 71–80.
[18] G. Arminger, C.C. Clogg, M.E. Sobel (Eds.), Handbook of Statistical Modeling for the Social and Behavioral Sciences, Plenum Press, New York, 1995.
[19] T. Hastie, R. Tibshirani, Generalized Additive Models, Chapman and Hall, New York, 1997.
[20] C.M. Bishop, Pattern Recognition and Machine Learning, Springer, New York, 2006.
[21] S. Lipovetsky, PCA and SVD with nonnegative loadings, Pattern Recognition 42 (2009) 68–76.
[22] S. Lipovetsky, Post-stratification with optimized effective base: Linear and nonlinear ridge regression approach, in: Proceedings of the Joint Statistical Meeting, American Statistical Association, Salt Lake City, Utah, July–August 2007, pp. 2313–2320.
[23] S. Becker, Y. Le Cun, Improving the convergence of back-propagation learning with second order methods, in: D.S. Touretzky, G.E. Hinton, T.J. Sejnowski (Eds.), Proceedings of the 1988 Connectionist Models Summer School, Morgan Kaufmann, San Mateo, CA, 1988, pp. 29–37.
[24] E.A. Bender, Mathematical Methods in Artificial Intelligence, IEEE Computer Society Press, Los Alamitos, CA, 2000.
[25] C.R. Rao, Linear Statistical Inference and its Application, Wiley, New York, 1973.
[26] A.S.C. Ehrenberg, How good is best, Journal of Royal Statistical Society A145 (1982) 364–366.
[27] S. Lipovetsky, Two-parameter ridge regression and its convergence to the eventual pairwise model, Mathematical and Computer Modelling 44 (2006) 304–318.
[28] P. McCullagh, J.A. Nelder, Generalized Linear Models, Chapman and Hall, New York, 1997.
[29] J.M. Chambers, T.J. Hastie, Statistical Models in S, Wadsworth and Brooks, Pacific Grove, CA, 1992.
[30] S-Plus'2000, MathSoft, Seattle, WA, 1999.