Journal of Modern Applied Statistical Methods

Volume 11 | Issue 2 Article 4

11-1-2012

Regression Split by Levels of the Dependent Variable

Stan Lipovetsky
GfK Custom Research North America, Minneapolis, MN, stan.lipovetsky@gfk.com

Journal of Modern Applied Statistical Methods
November 2012, Vol. 11, No. 2, 319-324

1538 – 9472/12/$95.00

Regression Split by Levels of the Dependent Variable

Stan Lipovetsky
GfK Custom Research North America,
Minneapolis, MN

Multiple regression coefficients split by the levels of the dependent variable are examined. The decomposition of the coefficients can be defined by points on the ordinal scale or by levels in the numerical response using the Gifi system of binary variables. This approach permits consideration of specific values of the coefficients at each layer of the response variable. Numerical results illustrate how to identify levels of interpretable regression coefficients.

Key words: Regression model, Gifi system, regression coefficients, levels of response.

Introduction

Interpretation of ordinary least squares (OLS) multiple linear regression with multicollinearity is a well-known problem that has been described in numerous works. The problem is caused by a deteriorating effect that multicollinearity between predictors can produce on OLS coefficients. OLS yields the best aggregate of predictors to fit data – and it is perfect for prediction – but it was not designed to obtain meaningful coefficients for individual predictors in regression (Abraham & Ledolter, 1983; Weisberg, 1985; Andersen & Skovgaard, 2010). Depending on the data, such a model could be useless in analyzing predictor impact on the dependent variable (DV) because multicollinearity yields inflated regression coefficients, pushing them towards large values of both signs. For example, if multicollinearity yields a negative sign for a presumably useful variable in the model, it is difficult to decide whether it makes sense to increase the value of such a variable to obtain a lift in the output. The techniques for constructing regression models with interpretable coefficients and contributions to the explained variance include ridge regressions (Hoerl & Kennard, 1970, 2000) and various other techniques, particularly: Shapley value regression, logit and multinomial parameterization of coefficients, and models by data gradients (Lipovetsky & Conklin, 2001, 2010c; Lipovetsky, 2009, 2010a, b; Nowakowska, 2010).

The possibility of splitting regression coefficients by the levels of the response variable and studying them separately is considered herein. This will help identify how the obtained coefficients are composed depending on the different values reached by the dependent variable (DV), and how this composition creates the total values of the coefficients. The technique is demonstrated for a DV measured using a numerical and rating scale (such as a Likert-type scale from 1 to 5 or 1 to 10), using the so-called Gifi system of binary multivariables (Gifi, 1990; Michailidis & de Leeuw, 1998; Mair & de Leeuw, 2010), where a variable on a several-point scale can be represented as a set of binary variables – one for each level. For example, a DV on a 5-point scale is presented as the first binary variable with ones in the place of 1s in the original variable and zeroes otherwise, up to the fifth binary variable

Stan Lipovetsky, Ph.D. is Senior Research Director at the GfK Research Center for Excellence, Marketing Sciences. He has numerous publications in multivariate statistics, multiple criteria decision making, econometrics, microeconomics and marketing research. He is a member of the editorial boards of the International Journal of Operations and Quantitative Management, the Journal of Electronic Modeling and the Journal of Model Assisted Statistics and Applications. Email him at: stan.lipovetsky@gfk.com.

where ones represent 5s in the original variable and zeroes otherwise. It is sometimes convenient to consider a fewer number of binary variables, for example, in key dissatisfaction analysis (Conklin, et al., 2004), it is sufficient to use only three binary variables: dissatisfaction (lower levels), neutral (middle), and enhanced values (upper levels). Regressions with interpretable coefficients attained by the split solutions help decision makers and managers understand the results of statistical modeling.

Regression Coefficients by Levels of the DV

A multiple linear regression can be presented as the model

$$y_i = a_0 x_{i0} + a_1 x_{i1} + a_2 x_{i2} + ... + a_n x_{in} + \varepsilon_i \quad (1)$$

where y_i and x_{ij} are i^{th} observations (i = 1, ..., N) by the DV y and by each j^{th} independent variable x_j (j = 0, 1, 2, ..., n), a_j are coefficients of the regression, including the intercept a_0 related to the identity variable x_0, and ε_i denotes added random noise. The OLS objective minimizes the squared errors ε_i and yields the solution which, in matrix notation, is:

$$\mathrm{a} = (\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathrm{y} \quad (2)$$

where a denotes the vector of all coefficients of regression (1), X is the design matrix of N by 1+n order of all the predictors, prime denotes transposition and vector y is of the N^{th} order.

Formula (2) shows that the regression coefficients are linear combinations of the y values aggregated with the coefficients of the transfer operator $T \equiv (X'X)^{-1}X'$ which depends only on the independent variables. Each j^{th} coefficient a_j is defined as a scalar product of the vector y and the values in the j^{th} row of this matrix T. Therefore, if the vector y is presented as a sum of several sub-vectors then it is possible to obtain the coefficients (2) related to each of these components.

Suppose y is measured in a rating scale of K values, so it can be presented as:

$$y = m_1 d_1 + m_2 d_2 + ... + m_K d_K, \quad (3)$$

where each d_k (k = 1, 2, …, K) is a binary vector of the N^{th} order, which has ones in the positions where y_i has the value k, otherwise it consists of zeros. For example, if y_i = 3 for i = 10, 15 and 18, then the binary vector d_3 has ones in the same 10^{th}, 15^{th} and 18^{th} places, otherwise zero, and similarly with the other vectors. Such a system of binary variables is called the Gifi system. The constant coefficients m_k in (3) for a Likert scale with ratings from 1 to K coincide with these values, so $m_k = k$. If y is a numerical variable, then it can be divided into several segments by its increasing values, and coefficients m_k represent the mean y values within each segment while the Gifi binary vectors d_k show by 1 and 0 values the particular segment to which each y_i belongs.

Substituting (3) into (2) yields the decomposition of the regression coefficients by the levels of y:

$$\begin{aligned} \mathrm{a} &= (\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\left(\mathrm{m}_1\mathrm{d}_1 + \mathrm{m}_2\mathrm{d}_2 + ... + \mathrm{m}_K\mathrm{d}_K\right) \\ &= \mathrm{m}_1(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathrm{d}_1 + \mathrm{m}_2(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathrm{d}_2 \\ &\quad + ... + \mathrm{m}_K(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathrm{d}_K \end{aligned} \quad (4)$$

Each matrix product $(X'X)^{-1}X'd_k$ in (4) is the vector of regression coefficients of the binary variable d_k by all the predictors x. It can also be described as the Fisher discriminator of the observations' assignment to each k^{th} segment of the data, thus the total regression coefficients are presented as the linear combination of these discriminators. For each particular level of y-values the coefficients of the Gifi response regressions can be denoted as:

$$\mathrm{b}^{(\mathrm{k})} \equiv (\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathrm{d}_\mathrm{k}, \quad (5)$$

and the items in decomposition (4) of the total vector of regression coefficients by the coefficients defined on each y-level are:

$$\begin{aligned} a^{(k)} &= m_k b^{(k)} \\ &= m_k (X'X)^{-1} X' d_k \end{aligned} . \quad (6)$$

It is also possible to consider regression results split by the independent variables, but the analysis becomes more complicated (Lipovetsky & Conklin, 2005).

Regression coefficients (2) or (4) can be presented as a sum of the by-level coefficients (6):

$$a = a^{(1)} + a^{(2)} + ... + a^{(K)} . \quad (7)$$

This result is very useful for practical applications of regression modeling because it permits consideration of the increments which can be reached specifically at any level of the dependent variable. For example, in marketing research studies on satisfaction with a product or service, it is useful to consider the regression subsets related to the lower levels of dissatisfaction and upper levels of enhanced satisfaction. The cumulative subtotal of coefficients can also be obtained by adding the needed vectors (7): its application to the multicollinearity problem is discussed next using numerical examples.

Numerical Examples

Consider a marketing research project with 623 respondents evaluating their satisfaction with a bank on a Likert scale from 1 (worst value) to 5 (best value). The variables are: y, overall satisfaction; x_1, customer service regarding checking account; x_2, explanations of features; x_3, kept informed of changes; x_4, convenient branch locations; x_5, convenient ATM locations; x_6, error free checking; x_7, representative solves problems; and x_8, clear comprehensive statements. The constructed OLS regression model (1)-(2) is:

$$\begin{aligned} y = {} & 1.664 + 0.155x_1 + 0.150x_2 \\ & + 0.123x_3 + 0.005x_4 - 0.012x_5 \\ & + 0.044x_6 + 0.137x_7 - 0.012x_8 \end{aligned} . \quad (8)$$

The main impact on overall satisfaction comes from predictors x_1, x_2, x_3 and x_7. Despite the positive impact that can be assumed for each driver on satisfaction, which is supported by positive pair correlations of each x with y, multicollinearity makes x_4 and x_6 negligibly small coefficients and yields a negative influence on both x_5 and x_8.

The five Gifi binary regressions for each level of overall satisfaction estimated by (5) are presented in the columns of Table 1. It is evident that the predictors have mixed coefficients for all levels of y, except the top level (k = 5), which has all positive coefficients of regression (the negative is the intercept).

Multiplying vectors (5) in the columns of Table 1 by the values $m_k = k$ transforms them into the components (6) of the original regression. These coefficients (6) and their total (7) are shown in Table 2. The coefficients of the last column in Table 2 coincide with the coefficients of the OLS model (8). In contrast to these OLS coefficients with both signs, the k = 5 model yields all positive coefficients.

Another useful way to consider the regression coefficients by splitting the cumulative levels of the response is shown in Table 3. It is clear from relation (7) that it is possible to consider subtotals of the split to lower and upper levels. Table 3 presents pairs of the models of the first level versus all other levels (columns denoted as 1 vs. 2:5), two lower and three upper levels (columns 1:2 vs. 3:5), three lower and two upper levels (columns 1:3 vs. 4:5), then four lower versus one upper level (columns 1:4 vs. 5), and finally the total regression by all the levels together (1)-(2). The last row in Table 3 presents the coefficients of multiple determination, R^2, well-known as a convenient characteristic of quality of the regression model. The sum of the coefficients in each pair of lower and upper models yields the total OLS coefficients of the last column in Table 3. This is not true with the coefficient of multiple determination, R^2, which is not a linear function of the DV values.

It is observed that a sum of two R^2 values in the paired columns in Table 3 can be higher or lower but not equal to R^2=0.297 of the total model. Table 3 also shows that the expected signs of the predictors' relation to the

dependent variable are given only by the upper level of overall satisfaction. Similar results are observed in various data sets.

Conclusion

Decomposition of multiple regression coefficients by the levels of the dependent variable was considered using the Gifi system of binary variables. The coefficients' split by the levels of the response variable can be easily performed with any software for ordinary least squares regression. The results of this decomposition help identify the subsets of the coefficients not distorted by multicollinearity and find an adequate interpretation of the regression coefficients that will be useful for managerial decisions.

References

Abraham, B., & Ledolter, J. (1983). *Statistical methods for forecasting*. New York, NY: Wiley.

Andersen, P. K., & Skovgaard, L. T. (2010). *Regression with linear predictors*. New York, NY: Springer.

Conklin, M., Powaga, K., & Lipovetsky, S. (2004). Customer satisfaction analysis: Identification of key drivers. *European Journal of Operational Research*, *154*, 819-827.

Gifi, A. (1990). *Nonlinear multivariate analysis*. Chichester, England: Wiley.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for non-orthogonal problems. *Technometrics*, *12*, 55-67.

Table 1: Coefficients of Gifi Response Regressions (5)

Coefficients	$b^{(1)}$	$b^{(2)}$	$b^{(3)}$	$b^{(4)}$	$b^{(5)}$
$b_0^{(k)}$	0.156	0.175	0.729	0.727	-0.787
$b_1^{(k)}$	-0.008	-0.030	-0.017	0.003	0.052
$b_2^{(k)}$	-0.009	-0.027	0.012	-0.054	0.079
$b_3^{(k)}$	-0.012	0.000	-0.046	0.018	0.040
$b_4^{(k)}$	0.010	0.004	-0.020	-0.015	0.021
$b_5^{(k)}$	0.006	0.014	-0.012	-0.029	0.021
$b_6^{(k)}$	-0.006	0.007	-0.036	0.031	0.004
$b_7^{(k)}$	-0.017	-0.014	-0.027	0.024	0.033
$b_8^{(k)}$	0.005	0.010	0.005	-0.051	0.030

Table 2: Coefficients of Regression Split by Levels of Response (6)

Coefficients	$a^{(1)}$	$a^{(2)}$	$a^{(3)}$	$a^{(4)}$	$a^{(5)}$	Total a
$a_0^{(k)}$	0.156	0.350	2.188	2.907	-3.937	1.664
$a_1^{(k)}$	-0.008	-0.061	-0.051	0.013	0.261	0.155
$a_2^{(k)}$	-0.009	-0.054	0.035	-0.216	0.396	0.150
$a_3^{(k)}$	-0.012	0.000	-0.139	0.074	0.200	0.123
$a_4^{(k)}$	0.010	0.008	-0.061	-0.059	0.107	0.005
$a_5^{(k)}$	0.006	0.028	-0.037	-0.115	0.106	-0.012
$a_6^{(k)}$	-0.006	0.015	-0.108	0.125	0.018	0.044
$a_7^{(k)}$	-0.017	-0.027	-0.082	0.096	0.167	0.137
$a_8^{(k)}$	0.005	0.021	0.016	-0.205	0.151	-0.012

Table 3: Coefficients of Regression with the DV Split into Two Levels

	1 vs.2:5		1:2 vs. 3:5		1:3 vs. 4:5		1:4 vs. 5		1:5
	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Total
a_0	0.156	1.508	0.506	1.158	2.694	-1.030	5.602	-3.937	1.664
a_1	-0.008	0.163	-0.069	0.224	-0.120	0.275	-0.106	0.261	0.155
a_2	-0.009	0.160	-0.064	0.214	-0.029	0.180	-0.245	0.396	0.150
a_3	-0.012	0.136	-0.012	0.135	-0.151	0.274	-0.077	0.200	0.123
a_4	0.010	-0.005	0.017	-0.013	-0.043	0.048	-0.102	0.107	0.005
a_5	0.006	-0.018	0.034	-0.046	-0.003	-0.009	-0.118	0.106	-0.012
a_6	-0.006	0.050	0.008	0.035	-0.100	0.143	0.026	0.018	0.044
a_7	-0.017	0.154	-0.044	0.181	-0.126	0.263	-0.030	0.167	0.137
a_8	0.005	-0.017	0.026	-0.037	0.042	-0.053	-0.163	0.151	-0.012
R^2	0.070	0.289	0.139	0.288	0.184	0.280	0.090	0.166	0.297

Hoerl, A. E., & Kennard, R. W. (2000). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *42*, 80-86.

Lipovetsky, S., & Conklin, M. (2001). Analysis of regression in game theory approach. *Applied Stochastic Models in Business and Industry*, *17*, 319-330.

Lipovetsky, S., & Conklin, M. (2005). Regression by data segments via discriminant analysis. *Journal of Modern Applied Statistical Methods*, *4*, 63-74.

Lipovetsky, S. (2009). Linear regression with special coefficient features attained via parameterization in exponential, logistic, and multinomial-logit forms. *Mathematical and Computer Modelling*, *49*, 1427-1435.

Lipovetsky, S. (2010a). Enhanced ridge regressions. *Mathematical and Computer Modelling*, *51*, 338-348.

Lipovetsky, S. (2010b). Meaningful regression coefficients built by data gradients. *Advances in Adaptive Data Analysis*, *2*, 451-462.

Lipovetsky, S., & Conklin, M. (2010c) Meaningful regression analysis in adjusted coefficients Shapley value model. *Model Assisted Statistics and Applications*, *5*, 251-264.

Mair, P., & de Leeuw, J. (2010). A general framework for multivariate analysis with optimal scaling: The R package aspect. *Journal of Statistical Software*, *32*, 1-23.

Michailidis, G., & de Leeuw, J. (1998). The Gifi system of descriptive multivariate analysis. Statistical Science 13: 307-336.

Nowakowska, E. (2010). Modeling in a multicollinear setup: Determinants of SVR advantage. *Model Assisted Statistics and Applications*, *5*, 219-233.

Weisberg, S. (1985). *Applied linear regression*. New York, NY: Wiley.