Received: 3 November 2023 | Revised: 25 December 2023 | Accepted: 7 January 2024 | Published online: 8 January 2024

RESEARCH ARTICLE

Journal of Data Science and Intelligent Systems
2024, Vol. 00(00) 1–7
DOI: 10.47852/bonviewJDSIS42021995

BON VIEW PUBLISHING

Multiple Regression Model as Interpolation Through the Points of Weighted Means

Stan Lipovetsky[1],*

[1]*Independent Consultant, USA*

Abstract: A well-known property of the multiple linear regression is that its plane goes through the point of the mean values of all variables, and this feature can be used to find the model's intercept. This work shows that a regression by n predictors also passes via n additional points of the specific weighted mean values. Thus, the regression is uniquely defined by all these $n+1$ multidimensional points of means, and approximation of observations by the theoretical model collapses to the interpolation function going through the knots of the weighted means. This property is obtained from the normal system of equations which serves for finding the linear regression parameters in the ordinary least squares approach. The derived features can be applied in nonlinear modeling for adjusting the model parameters so that the fitted values would go through the same reference points of means, that can be useful in applied regression analysis. Numerical examples are discussed. The found properties reveal the essence of regression function as hyperplane going through special points of mean values, which makes regression models more transparent and useful for solving and interpretation in various applied statistical problems.

Keywords: multiple regression, normal system of equations, weighted mean values, nonlinear modeling

1. Introduction

A new interpretation of regression model as interpolation through several special points of weighted mean values is suggested. It is useful for explaining the linear model features, and for adjustment in the nonlinear models.

One of the main tools of the applied statistics used in various estimations by the observed data is linear regression modeling. Such models can be built by minimization of the total squared errors by the response variable, that produces the so-called normal system of equations for estimation of the model parameters. This model is known as the ordinary least squares (OLS) regression which helps in the analysis, classification, prediction, finding key drivers among the explanatory variables, and many other aims of practical statistical evaluations.

There are tons of books and articles devoted to the OLS models and their applications (Andersen & Skovgaard, 2010; Cook, 2009; Draper & Smith, 1998; Gentle, 2017; Grafarend & Awange, 2012; Hilbe & Robinson, 2013; Hocking, 2013; Kendall & Stuart, 1967; Kuhn & Johnson, 2013; Matloff, 2017; Montgomery et al., 2021; Weisberg, 2005; Young, 2018). Multiple works describe software on the topics of linear and nonlinear statistical modeling (Demidenko, 2019; Efron & Hastie, 2021; Faraway, 2021; Irizarry, 2019; James et al., 2013; Venables & Ripley, 1999).

The current paper describes a hardly known property of the OLS regression: its line, or plane (or hyperplane in the case of multiple predictors) always goes via several points defined by some specific weighted mean values. A much more known property of the OLS linear regression is that it passes through the point of the mean values of its variables, and this feature is used in finding the model's intercept. The current work extends this property and shows that a regression by n predictors also goes through n other points of weighted mean values. These mean values can be calculated from the data even before the regression is built. Thus, the approximation of observations by the theoretical model with $1+n$ parameters (the intercept and coefficients of regression) can be considered as the interpolation through all the knots defined by the weighted means. This property is obtained from the normal system of equations which is commonly used for finding the linear regression parameters. Consideration of such properties was started by Lipovetsky and Conklin (2001), and the current work develops this study further.

The mentioned knots of the special weighted mean values can be called the reference points. The newly introduced reference points play the main role in reducing a regression approximation to the model of interpolation. As shown in the current work, approximation of the data by the theoretical model in the OLS approach can be seen as collapsing observations to the special weighted means used for the interpolation function. This approach reminds some other techniques of reduction observations to their frequencies, or clusters, helpful to build both linear and nonlinear models (Lipovetsky, 2015; Lipovetsky, 2019).

The same reference points property can be applied in nonlinear modeling, especially for the generalized linear models (GLMs) which can be represented via the linear-link functions (McCullagh & Nelder, 1989). In some problems, it is required to adjust GLM parameters so that the fitted values would go through the same points of the original mean values. In the work of King and Zeng (2001), it was considered how to modify an intercept in logistic regression so that it would better reproduce the original frequency of the binary response. The current work shows how to build a logit model which fits the reference points of the original weighted mean values, and predicts the probability within the required 0–1 interval.

*Corresponding author: Stan Lipovetsky, Independent Consultant, USA. Email: stan.lipovetsky@gmail.com

Numerical examples show that this approach can help in regression interpretation and in solving various applied statistical problems.

The paper is structured as follows. After introduction, Section 2 describes the simple linear regression as a line connecting two points. Section 3 considers the OLS normal equations and their transformation to the problem of interpolation by several reference points of weighted means. Section 4 extends this technique to the GLM linear link-function. Section 5 presents numerical examples, and Section 6 summarizes.

2. Simple Linear Regression as a Line Through Two Reference Points

Let us consider the linear regression in a simple case of two observed variables x and y, when the model for the dependent variable can be written as

$$y_i = a_0 + a_1 x_i + \varepsilon_i \tag{1}$$

in which i denotes observations ($i = 1, 2, \ldots, N$, where N – their total number), a_0 and a_1 are parameters of the model, and ε is the error by y. The OLS criterion consists in minimizing the sum of squared deviations:

$$S^2 = \sum_{i=1}^N \varepsilon_i^2 = \sum_{i=1}^N (y_i - a_0 - a_1 x_i)^2 \tag{2}$$

Equalizing the derivatives by the unknown coefficients to zero yields the so-called normal system:

$$\begin{cases} \sum_{i=1}^N y_i = a_0 N + a_1 \sum_{i=1}^N x_i \\ \sum_{i=1}^N x_i y_i = a_0 \sum_{i=1}^N x_i + a_1 \sum_{i=1}^N x_i^2 \end{cases} \tag{3}$$

Solving the system of linear by the parameters a_0 and a_1 Equations (3) produces the estimates for the regression coefficients – the intercept and slope:

$$a_0 = \overline{y} - a_1 \overline{x}\,, \qquad a_1 = \frac{covar(x, y)}{var(x)} = \frac{\sum_{i=1}^N (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^N (x_i - \overline{x})^2} \tag{4}$$

in which the bar above a variable denotes its mean value. The slope a_1 is defined as the quotient of the sample covariance of x and y to the variance of the independent variable x. The intercept a_0 in Equation (4) corresponds to the first equation in Equation (3) divided by the number of observations N, so the regression model in explicit form is

$$y = a_0 + a_1 x = \overline{y} + a_1 (x - \overline{x}) \tag{5}$$

in which the slope a_1 is defined in Equation (4). If the variable x equals its mean value, $x = \overline{x}$, then the variable y becomes equal its mean value, $y = \overline{y}$. Thus, in the plane with the coordinates x and y, the line Equation (5) of the linear regression goes through the point $(\overline{x}, \overline{y})$ of the mean values. As is known, a unique line in a plane always goes via two points, so what exactly is the second point for the Equation (5)?

To find it, let us assume that the total of x differs from zero, so we can divide the second equation Equation (4) by it and present the intercept a_0 as:

$$a_0 = \frac{\sum_{i=1}^N y_i x_i}{\sum_{i=1}^N x_i} - a_1 \frac{\sum_{i=1}^N x_i^2}{\sum_{i=1}^N x_i} = \sum_{i=1}^N \frac{x_i}{\sum_{j=1}^N x_j} y_i - a_1 \sum_{i=1}^N \frac{x_i}{\sum_{j=1}^N x_j} x_i \tag{6}$$

Introducing the weights of x values in their total,

$$w_i = \frac{x_i}{\sum_{j=1}^N x_j}, \qquad \sum_{i=1}^N w_i = 1\,, \tag{7}$$

it is possible to reduce the expression Equation (6) to the following:

$$a_0 = \sum_{i=1}^N w_i y_i - a_1 \sum_{i=1}^N w_i x_i = \overline{y_w} - a_1 \overline{x_w} \tag{8}$$

in which $\overline{y_w}$ and $\overline{x_w}$ are the weighted mean values of the y and x variables, respectively, with the weights of the x values in their total Equation (7). Therefore, the regression model, besides the point of mean values Equation (5), goes also through the second point of the weighted mean values $(\overline{x_w},\ \overline{y_w})$.

The line of pair regression going through two found points of the means $(\overline{x}, \overline{y})$ and the weighted means $(\overline{x_w},\ \overline{y_w})$ is defined by the first equation in Equations (4) and (8) for the intercept, that can be represented in the transformed system Equation (3) as follows:

$$\begin{cases} \overline{y} = a_0 + a_1 \overline{x} \\ \overline{y_w} = a_0 + a_1 \overline{x_w} \end{cases} \tag{9}$$

Solving this system for the coefficients of regression reveals the formulae

$$a_1 = \frac{\overline{y_w} - \overline{y}}{\overline{x_w} - \overline{x}}\,, \qquad a_0 = \overline{y} - \frac{\overline{y_w} - \overline{y}}{\overline{x_w} - \overline{x}} \overline{x} \tag{10}$$

which are the slope and intercept of the model. Then the regression Equation (5) expressed via the points of means and weighted means of the variables is

$$y = a_0 + a_1 x = \overline{y} + \frac{\overline{y_w} - \overline{y}}{\overline{x_w} - \overline{x}} (x - \overline{x}) = \frac{\overline{y}\,\overline{x_w} - \overline{y_w}\,\overline{x}}{\overline{x_w} - \overline{x}} + \frac{\overline{y_w} - \overline{y}}{\overline{x_w} - \overline{x}} x \tag{11}$$

Indeed, the equality $x = \overline{x}$ yields $y = \overline{y}$, and equality $x = \overline{x_w}$ leads to $y = \overline{y_w}$. It is easy to show that the slope in Equation (11) can be transformed to the expression

$$a_1 = \frac{\overline{y_w} - \overline{y}}{\overline{x_w} - \overline{x}} = \frac{\frac{\sum_{i=1}^N y_i x_i}{\sum_{i=1}^N x_i} - \overline{y}}{\frac{\sum_{i=1}^N x_i^2}{\sum_{i=1}^N x_i} - \overline{x}} = \frac{\sum_{i=1}^N y_i x_i - N \overline{y}\,\overline{x}}{\sum_{i=1}^N x_i^2 - N \overline{x}^2} = \frac{\sum_{i=1}^N (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^N (x_i - \overline{x})^2} \tag{12}$$

It coincides with the regular definition of the slope Equation (4), but the slope in the formulae (10)–(11) has the explicit meaning of the line going through two points of the mean and weighted mean values.

3. Multiple Linear Regression as Interpolation by Reference Points

Let us extend the above given derivation to the general case of n predictors, when the linear model for the dependent variable is

$$y_i = a_0 + a_1 x_{1i} + a_2 x_{2i} + \ldots + a_n x_{ni} + \varepsilon_i \tag{13}$$

and the OLS criterion can be presented as follows:

$$S^2 = \sum_{i=1}^N \varepsilon_i^2 = \sum_{i=1}^N (y_i - a_0 - a_1 x_{1i} - \ldots - a_n x_{ni})^2. \tag{14}$$

Minimization of the objective Equation (14) consists in taking derivatives by the unknown coefficients and equalizing them to

zero, which yields the normal system of equations:

$$\begin{cases} \sum_{i=1}^{N} y_i = a_0 N + a_1 \sum_{i=1}^{N} x_{1i} + \ldots + a_n \sum_{i=1}^{N} x_{ni} \\ \sum_{i=1}^{N} x_{1i} y_i = a_0 \sum_{i=1}^{N} x_{1i} + a_1 \sum_{i=1}^{N} {x_{1i}}^2 + \ldots + a_n \sum_{i=1}^{N} x_{1i} x_{ni} \\ - \quad - \quad - \quad - \quad - \quad - \quad - \quad - \quad - \quad - \quad - \quad - \\ \sum_{i=1}^{N} x_{ni} y_i = a_0 \sum_{i=1}^{N} x_{ni} + a_1 \sum_{i=1}^{N} x_{ni} x_{1i} + \ldots + a_n \sum_{i=1}^{N} {x_{ni}}^2 \end{cases} \tag{15}$$

The first Equation (15) produces the well-known expression for the intercept:

$$a_0 = \overline{y} - a_1 \overline{x_1} - \ldots - a_n \overline{x_n} \tag{16}$$

which means that the hyperplane of the multiple linear regression goes via the point of mean values in the space of all the variables. Substituting the intercept Equation (16) into the other n Equations (15) reduces them to the following equation in the matrix form and its solution:

$$C_{xx} a = c_{xy}, \qquad a = C_{xx}^{-1} c_{xy} , \tag{17}$$

in which C_{xx} and c_{xy} are the n-th order sample covariance matrix and vector of the x variables among themselves, and with y variable, respectively. The solution for the vector a of the coefficients of regression is given via the inverted matrix C_{xx}^{-1}, which is convenient for calculations but is not helpful for understanding a meaning of the OLS result Equation (17).

To obtain a more interpretable form of the OLS solution, let us prove the following theorem: for the nonzero totals by each predictor, the approximation of observations by the multiple linear regression with $1+n$ parameters of the intercept and coefficients of regression can be presented as the hyperplane model of interpolation function through all $1+n$ points of the special weighted mean values.

To prove it, let us reformulate the normal system Equation (15) so that its solution would become clearly interpretable and meaningful. In assumption of the total of any x differs from zero, we divide each j-th Equation (15) by the term used in this equation with the intercept a_0, so the system Equation (15) becomes

$$\begin{cases} \frac{1}{N}\sum_{i=1}^{N} y_i = a_0 + a_1 \frac{1}{N}\sum_{i=1}^{N} x_{1i} + \ldots + a_n \frac{1}{N}\sum_{i=1}^{N} x_{ni} \\ \sum_{i=1}^{N} \frac{x_{1i}}{\sum_{j=1}^{N} x_{1j}} y_i = a_0 + a_1 \sum_{i=1}^{N} \frac{x_{1i}}{\sum_{j=1}^{N} x_{1j}} x_{1i} + \ldots + a_n \sum_{i=1}^{N} \frac{x_{1i}}{\sum_{j=1}^{N} x_{1j}} x_{ni} \\ - \quad - \quad - \quad - \quad - \quad - \quad - \quad - \quad - \quad - \quad - \quad - \quad - \\ \sum_{i=1}^{N} \frac{x_{ni}}{\sum_{j=1}^{N} x_{nj}} y_i = a_0 + a_1 \sum_{i=1}^{N} \frac{x_{ni}}{\sum_{j=1}^{N} x_{nj}} x_{1i} + \ldots + a_n \sum_{i=1}^{N} \frac{x_{ni}}{\sum_{j=1}^{N} x_{nj}} x_{ni} \end{cases} \tag{18}$$

The quotients in each j-th Equation (18) are nothing else but the weights of i-th observations in their total for each variable x_j. These sets of weights can be denoted as follows:

$$w_{1i} = \frac{x_{1i}}{\sum_{j=1}^{N} x_{1j}} , \quad \ldots , \quad w_{ni} = \frac{x_{ni}}{\sum_{j=1}^{N} x_{nj}} \tag{19}$$

with the totals for each set equals one:

$$\sum_{i=1}^{N} w_{1i} = 1 , \quad \ldots , \quad \sum_{i=1}^{N} w_{ni} = 1 \tag{20}$$

Using the weights Equation (19), we introduce notations for the weighted means for all values in Equation (18):

$$\begin{cases} \overline{y} = \frac{1}{N}\sum_{i=1}^{N} y_i , \quad \overline{x_1} = \frac{1}{N}\sum_{i=1}^{N} x_{1i} , \ \ldots \ , \quad \overline{x_n} = \frac{1}{N}\sum_{i=1}^{N} x_{ni} \\ \overline{y_{w1}} = \sum_{i=1}^{N} w_{1i} y_i, \quad \overline{x_{1.w1}} = \sum_{i=1}^{N} w_{1i} x_{1i}, \ldots , \quad \overline{x_{n.w1}} = \sum_{i=1}^{N} w_{1i} x_{ni} \\ - \quad - \quad - \quad - \quad - \quad - \quad - \quad - \quad - \quad - \quad - \quad - \quad - \\ \overline{y_{wn}} = \sum_{i=1}^{N} w_{ni} y_i, \quad \overline{x_{1.wn}} = \sum_{i=1}^{N} w_{ni} x_{1i}, \quad \ldots \ \overline{x_{n.wn}} = \sum_{i=1}^{N} w_{ni} x_{ni} \end{cases} \tag{21}$$

in which $\overline{y_{wj}}$ are the mean values of y, and $\overline{x_{k.wj}}$ are the mean values of x_k, weighted by the j-th set of the weights built by x_j ($j = 1, 2, \ldots, n$) in their total Equation (19). The system Equation (18) can be represented via the mean values Equation (21) in the equations:

$$\begin{cases} \overline{y} = a_0 + a_1 \overline{x_1} + \ldots + a_n \overline{x_n} \\ \overline{y_{w1}} = a_0 + a_1 \overline{x_{1.w1}} + \ldots + a_n \overline{x_{n.w1}} \\ - \quad - \quad - \quad - \quad - \quad - \quad - \quad - \\ \overline{y_{wn}} = a_0 + a_1 \overline{x_{1.wn}} + \ldots + a_n \overline{x_{n.wn}} \end{cases} \tag{22}$$

Each row in the expressions Equation (21) define one of the $1+n$ reference points of the mean values, and the system Equation (22) describes the hyperplane going through all these points in the $(1+n)$-dimensional space of all the variables, which proves the theorem.

Solving the system Equation (22), we obtain the same values of the parameters a of regression Equation (17), but they can be expressed via the coefficients of mean values used in this system. To relax the condition of the total by each predictor differs from zero, instead of the weighted mean values Equation (18) it is possible to work with the total of each variable weighted by the x values Equation (15), so dividing by the total of x values for normalization of the weights is not necessary. Though, for a clearer explanation and interpretation of the obtained results it is more convenient to employ the terms of the weighted mean values.

A convenient way to find the coefficients of the hyperplane going via the reference points Equation (22) is known in the analytic geometry (Korn & Korn, 2000). For this aim, the determinant equation based on the system Equation (22) can be employed:

$$\begin{vmatrix} y & x_0 & x_1 & \ldots & x_n \\ \overline{y} & 1 & \overline{x_1} & \ldots & \overline{x_n} \\ \overline{y_{w1}} & 1 & \overline{x_{1.w1}} & \ldots & \overline{x_{n.w1}} \\ - & - & - & - & - \\ \overline{y_{wn}} & 1 & \overline{x_{1.wn}} & \ldots & \overline{x_{n.wn}} \end{vmatrix} = 0 \tag{23}$$

The first row of this determinant consists of the variables' names, including x_0 for the variable identically equal one, $x_0 = 1$, which is used with the intercept. All the other rows contain numerical values of the coordinates of the reference points. If to take the first row's elements equal the elements of any other row, the determinant would have the same two rows, so it equals zero. Therefore, this hyperplane goes exactly via each of the reference points. Decomposing this determinant by the elements of the first row, yields the equation of the hyperplane in the explicit formula.

Let us take an example of the model with one predictor, when the Equation (23) becomes

$$\begin{vmatrix} y & x_0 & x_1 \\ \overline{y} & 1 & \overline{x_1} \\ \overline{y_{w1}} & 1 & \overline{x_{1.w1}} \end{vmatrix} = y \cdot \begin{vmatrix} 1 & \overline{x_1} \\ 1 & \overline{x_{1.w1}} \end{vmatrix} - x_0 \cdot \begin{vmatrix} \overline{y} & \overline{x_1} \\ \overline{y_{w1}} & \overline{x_{1.w1}} \end{vmatrix} + x_1 \cdot \begin{vmatrix} \overline{y} & 1 \\ \overline{y_{w1}} & 1 \end{vmatrix} = 0 \tag{24}$$

Recalling the identity $x_0 = 1$, and finding the determinants of the second order, we can solve the Equation (24) for the variable y, which yields:

$$y = \frac{\overline{y}\,\overline{x_{1.w1}} - \overline{y_{w1}}\,\overline{x_1}}{\overline{x_{1.w1}} - \overline{x_1}} + \frac{\overline{y_{w1}} - \overline{y}}{\overline{x_{1.w1}} - \overline{x_1}} x_1 \tag{25}$$

With only one predictor, we can simplify x_1 as x, and w_1 as w, so the relations Equations (19)–(20) reduce to the weights Equation (7). Then the formula (25) coincides with the expression (11) for the simple pair regression.

It is useful to note that if to work with the centered data as in the solutions Equations (4) and (17), so with the variables with zero mean values, then it is yet possible to find the reference points through which the plane or hyperplane of the models always go. The weights in that case are proportional to the items of the squared centered values in the total variance, and the weighted means are built by the partial slopes by observations. More detail on this and other weighting schemes are discussed in Lipovetsky and Conklin (2001).

4. Implementation for GLM Adjustment to the Reference Points

The reference points of the OLS linear regression can be used for adjustment of the nonlinear models. Sometimes it is needed to alter parameters of a nonlinear model so that the fitted values would go through the same points of the original mean values. It is especially useful for the generalized linear models, GLM, which can be represented via the linear-link functions. For example, it is possible to modify an intercept in logistic regression so that it would better reproduce the original frequency of the binary response, as it was shown in King and Zeng (2001). Let us describe how to build a modified logistic model which fits the reference points of the original weighted mean values, and at the same time, keeps the predicted probability values in the required 0–1 interval.

As it is well-known, for a binary outcome y, the logistic regression can be constructed by the maximum likelihood criterion, and the probability estimation p by this model is performed using the expression

$$p = \frac{1}{1+\exp(-(b_0+b_1x_1+\ldots+b_nx_n))} \tag{26}$$

in which b denotes coefficients with the predictors in the logit model. The expression Equation (26) can be rewritten in the linear-link function by the following transformation:

$$ln\frac{p}{1-p} = b_0 + b_1x_1 + \ldots + b_nx_n \tag{27}$$

Using the new dependent variable

$$z = ln\frac{p}{1-p} \tag{28}$$

the relation Equation (27) can be considered as a linear model. As it was described above in the relations Equations (21)–(23), multiple linear regression traverses the reference points of the mean and weighted mean values. Thus, if to require the logistic regression to pass through the same reference points, we define 1+n values of the probabilities p to be equal the mean and weighted mean values of the binary outcome variable y with the weights by the predictors' values Equations (19)–(20):

$$p_0 = \overline{y},\ \ p_1 = \overline{y_{w1}},\ \ \ldots,\ \ p_n = \overline{y_{wn}} \tag{29}$$

The values Equation (29) can be used for finding the dependent variable z Equation (28) for the left-hand side of the linear model Equation (27):

$$z_0 = ln\frac{\overline{y}}{1-\overline{y}},\quad z_1 = ln\frac{\overline{y_{w1}}}{1-\overline{y_{w1}}},\ \ldots,\quad z_n = ln\frac{\overline{y_{wn}}}{1-\overline{y_{wn}}} \tag{30}$$

Substituting the mean and weighted mean values of y variable in the Equations (21)–(23) by the set of these new variables Equation (30), we can solve this system of equations and obtain the estimation of the parameters of logistic regression in the linearized form Equation (27). Then we return to the initial logit model Equation (26) for making estimation of the probability at each point of observations. Due to the property of interpolation by the reference points, this logistic regression goes through the mean and weighted mean values Equation (29) of the original binary outcome variable. We can call this mix-model the OLS.Logit regression.

5. Numerical Example

For an explicit numerical illustration, a small dataset is taken with ten observations by two variables y and x, shown in Table 1. Additional columns contain the weights w of x values in their total and y and x values weighted. The last rows present needed for calculations the total values, the means, and the standard deviations (std).

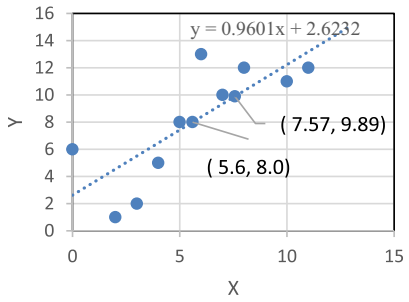
The coefficient of correlation between y and x is $r = 0.778$, so with the std values from Table 1 the regular regression estimation for the slope parameter (1)–(4) equals $a_1 = r*std(y)/std(x) = 0.778*4.32/3.5 = 0.96$. The intercept (4) equals $a_0 = \bar{y} - a_1\bar{x} = 8.0 - 0.96 * 5.6 = 2.62$. On the other hand, with the mean values $\bar{y} = 8.0$, $\bar{x} = 5.6$, and the weighted mean values $\overline{y_w} = 9.89$, $\overline{x_w} = 7.57$, the slope parameter (10) equals $a_1 = (9.89–8.0)/(7.57–5.6) = 0.96$. Thus, it coincides with the regular estimation, and the intercept (10) in the new estimation is 2.62, as expected.

Table 1
Example 1: A dataset with ten observations

i	y	x	w	$y*w$	$x*w$
1	1	2	0.04	0.04	0.07
2	6	0	0.00	0.00	0.00
3	2	3	0.05	0.11	0.16
4	10	7	0.13	1.25	0.88
5	8	5	0.09	0.71	0.45
6	12	11	0.20	2.36	2.16
7	11	10	0.18	1.96	1.79
8	12	8	0.14	1.71	1.14
9	13	6	0.11	1.39	0.64
10	5	4	0.07	0.36	0.29
Total	80	56	1	9.89	7.57
Mean	8.0	5.60	0.10	9.89	7.57
Std	4.32	3.50	0.06	0.86	0.74

The scatterplot and the line of regression for the data from Table 1 are presented in Figure 1. The regression equation and its coefficient of determination are $y = 0.96x + 2.62$ and $R^2 = 0.61$, so the quality of the model is good. The points of the mean values and the weighted mean values are also shown in the graph, and the interpolation line passing through these two points coincides, as we now know, with the linear regression trend.

Figure 1
Scatterplot, regression line, and the points of mean and weighted mean values



For the second example, the dataset "attitude" is taken from the R package "datasets". This dataset is considered in Chatterjee and Hadi (2000), and it describes a survey of the employees in a large financial organization, with the data aggregated from the questionnaires for respondents from thirty departments. The results of survey define the percent of favorable responses by the following variables: y – overall rating; x_1 – handling of employee complaints; x_2 – does not allow special privileges; x_3 – opportunity to learn; x_4 – raises based on performance; x_5 – too critical; x_6 – advancement. The data were transformed from the percent into the decimal portions, and for the aim of example the outcome y was made a binary variable with the value 0 or 1 when it was less or not of the mean level, respectively.

For building the OLS multiple regression, we find the normal system of Equations (15) which is given in Table 2.

Table 2
Example 2: Normal system for the OLS regression

	y	x_0	x_1	x_2	x_3	x_4	x_5	x_6
x_0	17.00	30.00	19.98	15.94	16.91	19.39	22.43	12.88
x_1	12.72	19.98	13.82	10.88	11.53	13.18	15.01	8.67
x_2	9.91	15.94	10.88	8.90	9.19	10.47	11.97	6.97
x_3	10.42	16.91	11.53	9.19	9.93	11.16	12.68	7.45
x_4	11.80	19.39	13.18	10.47	11.16	12.85	14.61	8.50
x_5	12.93	22.43	15.01	11.97	12.68	14.61	17.05	9.71
x_6	7.75	12.88	8.67	6.97	7.45	8.50	9.71	5.84

The first column in Table 2 presents the vector at the left-hand side and then goes the matrix at the right-hand side of the system (15). The column of the variable x_0 contains the totals of observations by each predictor, which corresponds to the coefficients used with the intercept in (15). Dividing the rows in Table 2 by these total values yields the normalized system (18) shown by its coefficients in Table 3.

Coefficients in Table 3 correspond to the mean and weighted mean values defining the reference points in the expressions (21)–(23) as well. Solving such a system yields the parameters of

Table 3
Example 2: Normal system for the OLS regression given in the reference points

	y	x_0	x_1	x_2	x_3	x_4	x_5	x_6
x_0	0.567	1	0.666	0.531	0.564	0.646	0.748	0.429
x_1	0.637	1	0.692	0.545	0.577	0.660	0.751	0.434
x_2	0.622	1	0.683	0.559	0.577	0.657	0.751	0.437
x_3	0.616	1	0.682	0.543	0.587	0.660	0.750	0.440
x_4	0.609	1	0.680	0.540	0.575	0.663	0.753	0.439
x_5	0.576	1	0.669	0.534	0.565	0.651	0.760	0.433
x_6	0.602	1	0.673	0.541	0.578	0.660	0.754	0.453

the Equation (13) estimated by the OLS regression modeling with help of the *lm* function in *R*.

These OLS parameters are shown in the first column of the numerical results in Table 4.

Table 4
Example 2: Parameters of OLS, Logit, and OLS.Logit regression models

	OLS	Logit	OLS.Logit
Intercept	−1.443	−24.473	−8.040
x_1	2.468	44.452	10.501
x_2	0.375	9.334	1.505
x_3	−0.056	−11.434	−0.301
x_4	−0.030	−9.627	−0.362
x_5	−0.098	−3.126	−0.437
x_6	0.678	14.037	2.900

The next column of Table 4 presents coefficients of the regular logit regression for the binary outcome, built by the known *glm R* function. The last column in Table 4 contains the adjustment of the logistical regression described in the formulae (26)–(30) for the OLS.Logit model which satisfies the reference points. This model parameters can be obtained using the function *solve* in *R*, with the matrix of the reference points (22) and vector of the dependent variable defined in Equation (30).

The three solutions in Table 4 could seem rather different; however, they are very highly correlated – their correlation matrix is shown in Table 5.

Table 5
Example 2: Correlations of vectors of OLS, Logit, and OLS.Logit solutions

Solution	OLS	Logit	OLS.Logit
OLS	1	0.972	0.996
Logit	0.972	1	0.958
OLS.Logit	0.996	0.958	1

Such a high correlation between different sets of the models' parameters leads to the predicted values close between themselves. For each of the three models considered in Table 4, the fitted values for the outcome variable y were estimated: by the OLS linear model Equation (13), by the logit model Equation (26), and

by the OLS.Logit Equations (26)–(30) regression. Table 6 shows correlations of the outcome y with the three fitted sets.

Table 6
Example 2: Correlations of the outcome y and fitted by OLS, Logit, and OLS.Logit values

Variable	y	OLS.fit	Logit.fit	OLS.Logit.fit
y	1	0.736	0.778	0.761
OLS.fit	0.736	1	0.928	0.989
Logit.fit	0.778	0.928	1	0.962
OLS.Logit.fit	0.761	0.989	0.962	1

All three vectors of fitted values are highly correlated and they are also in good correspondence with the observed y variable. The squared correlations of y with each of its fit vectors can serve as characteristics of multiple determination for the models presented in Table 4. It is useful to note that some predictions by the OLS model are below zero or above one, while both Logit and OLS.Logit produce all predictions within the meaningful probability interval from 0 to 1.

Dichotomizing the predicted variables as 0 or 1 for the values below the level 0.5 and values equal or higher than 0.5, we construct the cross-tables of counts for the original versus the predicted values – see Table 7.

The last row in Table 7 presents the hit rate, or the total of the correct predictions (sum of the diagonal counts) in the total number of observations. The OLS and OLS.Logit demonstrate a higher hit rate, but in general all models support results of each other.

The third example considers the physicochemical features of the red wine described in Cortez et al. (2009) with the data available at the repositories. From that dataset of 1599 observations by 12 variables, the three predictors mostly correlated with the outcome y variable were taken – those are: y – quality of wine (sensory preference in 10-point scale); x_1 – volatile acidity (g/dm^3); x_2 – sulphates (g/dm^3); and x_3 – alcohol (% vol).

The OLS multiple regression was built by solving the normal system of Equations (15). For these data, the second moments of the variables comprising the coefficients of the normal system are presented in Table 8. It is arranged similarly to Table 2, with the first column corresponding to the vector at the left-hand side Equation (15), and the next columns correspond to the matrix at the right-hand side of Equation (15). The column of x_0 contains the totals of the observations by each predictor.

Dividing rows in Table 8 by these totals by each predictor, we obtain the normalized system Equation (18), whose coefficients are shown in Table 9.

Coefficients in Table 9 present the mean and weighted mean values of the reference points in the expressions Equations (21)–(23). Solving any of these systems, we obtain the parameters of the Equation (13).

The pair correlations of y with xs and results of the OLS regression modeling performed in R are presented in Table 10.

We see that t-statistics for each coefficient of regression is much bigger than the threshold of $z = 1.96$ so all parameters are significant with high confidence probability. The coefficient of multiple determination R^2 in this model equals 0.336, with F-statistics equals 369, so the model is statistically significant.

In Cortez et al. (2009), it is discussed that a wine certification is usually assessed by the physicochemical and sensory tests, and the latter ones rely mainly on human experts. However, taste is the least understood of the human senses, and relationships between physicochemical observations and sensory tests are complex and still not fully understood. It makes wine classification a difficult task in which the regression analysis can help, allowing to obtain reliable predictions of the wine quality by the main predictors.

Table 7
Example 2: Cross-tables of original versus predicted counts by OLS, Logit, and OLS.Logit models

	OLS		Logit		OLS.Logit	
	Predic-ted $y = 0$	Predic-ted $y = 1$	Predic-ted $y = 0$	Predic-ted $y = 1$	Predic-ted $y = 0$	Predic-ted $y = 1$
Obser-ved $y = 0$	11	2	10	3	11	2
Obser-ved $y = 1$	3	14	3	14	3	14
Hit rate	0.833		0.800		0.833	

Table 8
Example 3: Normal system of equations

	y	$x0$	$x1$	$x2$	$x3$
$x0$	9012	1599	844	1052	16666
$x1$	4666	844	497	543	8735
$x2$	5986	1052	543	739	10996
$x3$	94587	16666	8735	10996	175528

Table 9
Example 3: Normal system of equations given in the reference points

	y	$x0$	$x1$	$x2$	$x3$
$x0$	5.636	1	0.528	0.658	10.423
$x1$	5.529	1	0.589	0.643	10.350
$x2$	5.688	1	0.516	0.702	10.449
$x3$	5.675	1	0.524	0.660	10.532

Table 10
Example 3: Correlation and results of the OLS regression modeling

	Cor(y, x)	Regression	Std	t-value
Intercept		2.611	0.196	13.34
$x1$	−0.39	−1.221	0.097	−12.59
$x2$	0.251	0.679	0.101	6.737
$x3$	0.476	0.309	0.016	19.57

6. Conclusion

The paper considered a useful feature of the multiple linear regression that its hyperplane always goes through several reference points defined by the special weighted mean values. It is an important enrichment of the well-known property that regression goes via the point of the mean values of variables. The current work extends this property and proves that all coefficients of regression can be found by the interpolation function passing through the reference points of the weighted mean values.

This property is also applied to the nonlinear modeling, especially to the generalized linear models, if an adjustment of parameters is needed to make the fitted values to pass through the points of the original mean values. Numerical examples show that this approach can be useful in regression interpretation and in solving various practical regression problems.

As a natural step in extension of the proposed approach to other methods of multivariate modeling, it is possible to indicate that the linear discriminant analysis (LDA), can be described as the regression of the dichotomic outcome by the predictors. For future research, the proposed approach of reference points can be tried for the LDA, and some other multivariate statistical techniques as well.

Recommendations

The found property of the multiple linear regression is very useful for teaching and learning of this main statistical tool because it clearly exposes its actual meaning commonly hidden behind the complicated formulae. It also permits to adjust the nonlinear models to the expected mean levels of the response variable.

Acknowledgments

The author is grateful to two referees whose comments and suggestions helped to improve the paper.

Ethical Statement

This study does not contain any studies with human or animal subjects performed by any of the authors.

Conflicts of Interest

The author declares that he has no conflicts of interest to this work.

Data Availability Statement

The data that support the findings of this study are openly available in [*Regression Analysis] by Example;* in [UC Irvine] at https://archive.ics.uci.edu/dataset/186/wine+quality; in [Kaggle] at https://www.kaggle.com/datasets/uciml/red-wine-quality-cortez-et-al-2009

References

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

Chatterjee, S., & Hadi, A. S. (2000). *Regression analysis by example*. USA: Wiley.

Cook, R. D. (2009). *Regression graphics: Ideas for studying regressions through graphics*. USA: John Wiley & Sons.

Cortez, P., Cerdeira, A., Almeida, F., Matos, T., & Reis, J. (2009). Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems*, *47*(4), 547–553.

Demidenko, E. (2019). *Advanced statistics with applications in R*. USA: John Wiley & Sons.

Draper, N. R., & Smith, H. (1998). *Applied regression analysis*. USA: John Wiley & Sons.

Efron, B., & Hastie, T. (2021). *Computer age statistical inference, student edition: Algorithms, evidence, and data Science*. UK: Cambridge University Press.

Faraway, J. J. (2021). *Linear models with python*. USA: CRC Press.

Gentle, J. E. (2017). *Matrix algebra: Theory, computations and applications in statistics*. Germany: Springer.

Grafarend, E. W., & Awange, J. L. (2012). *Applications of linear and nonlinear models: Fixed effects, random effects, and total least squares*. Germany: Springer.

Hilbe, J. M., & Robinson, A. P. (2013). *Methods of statistical model estimation*. USA: CRC Press.

Hocking, R. R. (2013). *Methods and applications of linear models: Regression and the analysis of variance*. USA: John Wiley & Sons.

Irizarry, R. A. (2019). *Introduction to data science: Data analysis and prediction algorithms with R*. USA: CRC Press.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning: With applications in R*. USA: Springer.

Kendall, M. G., & Stuart, A. (1967). *The advanced theory of statistics: Inference and relationship*. USA: Haffner Press.

King, G., & Zeng, L. (2001). Logistic regression in rare events data. *Political Analysis*, *9*, 137–163. https://doi.org/10.1093/oxfordjournals.pan.a004868

Korn, G. A., & Korn, T. M. (2000). *Mathematical handbook for scientists and engineers: Definitions, theorems, and formulas for reference and review*. USA: Dover Publications.

Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling*. USA: Springer.

Lipovetsky, S. (2015). Analytical closed-form solution for binary logit regression by categorical predictors. *Journal of Applied Statistics*, *42*(1), 37–49.

Lipovetsky, S. (2019). Regression modeling and prediction by individual observations versus frequency. *Journal of Modern Applied Statistical Methods*, *18*(1), 1.

Lipovetsky, S., & Conklin, M. (2001). Regression as weighted mean of partial lines: Interpretation, properties, and extensions. *International Journal of Mathematical Education in Science and Technology*, *32*(5), 697–706.

Matloff, N. (2017). *Statistical regression and classification: From linear models to machine learning*. USA: CRC Press.

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models*. USA: CRC Press.

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2021). *Introduction to linear regression analysis*. USA: John Wiley & Sons.

Venables, W. N., & Ripley, B. D. (1999). *Modern applied statistics with S-PLUS*. USA: Springer.

Weisberg, S. (2005). *Applied linear regression*. USA: John Wiley & Sons.

Young, D. S. (2018). *Handbook of regression methods*. USA: CRC Press.

How to Cite: Lipovetsky, S. (2024). Multiple Regression Model as Interpolation Through the Points of Weighted Means. *Journal of Data Science and Intelligent Systems*. https://doi.org/10.47852/bonviewJDSIS42021995