Multiple Regression: Methodology and Applications

Yiming Sun [1, †], Xinyuan Wang [1, †], Chi Zhang [2, †], Mingkai Zuo [3, *, †]

[1] Pennon Education, Qingdao, Shandong, China

[2] School of Science, Wuhan University of Technology, Wuhan, Hubei, China

[3] School of Statistics and Mathematics, Zhongnan University of Economics and Law, Wuhan, Hubei, China

* Corresponding author. Email: 201921090236@stu.zuel.edu.cn

[†]These authors contributed equally.

Abstract. Multiple regression is one of the most significant forms of regression and has a wide range of applications. The study of the implementation of multiple regression analysis in different settings contributes to the development of relevant theories and the improvement of models. In this paper, four different kinds of regressions are discussed individually by referring to different articles. The four kinds of regressions discussed are multivariable/multiple linear regression, multivariate multiple linear regression, multinomial logistic regression, and multivariate non-linear regression. As for multivariable/multiple linear regression, examples in the manufacturing industry and medical field show that it can be applied in more fields. Multivariate multiple linear regression is more accurate than multivariable/multiple linear regression and can be used with more than a variable. Multinomial logistic regression is relatively mature and accurate, and can help people well solve the problem with non-linearity and multiple independent variables. It does not require the variables to obey a multivariate normal distribution, and is more widely used as well. Multivariate non-linear regression, however, cannot be used properly without powerful professional knowledge. This paper investigates the theoretical development and model applications of multiple regression to demonstrate the flexibility and broadness of the adoption of multiple regression analysis.

Keywords: Multiple linear regression, Multivariate multiple linear regression, Multinomial logistic regression, Multivariate non-linear regression.

1. Introduction

Since the twentieth century, regression analysis approaches have been developed and expanded in many forms in order to be applied flexibly to a wide range of fields. Particularly, multiple regression is a critically valuable part of the analysis development process.

In 1931 [1], the multiple correlation and regression coefficients were modified by Bakst, through connecting with another method. Then, in 1938 [2], Bartlett carried out deeper research into the theory of multiple regression to reveal the further foundation of logical reasoning. Later, multiple regression analysis was extensively applied in the many various areas, such as the behavioural research and biology field [3, 4].

With the in-depth study of multiple regression analysis, multiple regression has also been adapted to different formats, including multivariate linear/non-linear regression and multinomial logistic regression, and transformed into new models to solve various types of issues. In 1975 [5], Maxwell explained the limited use of multiple regression model, while in 1980 [6], Hosmer transformed the model and gave the advantage of multiple logistic regression model for tests. However, the methods of multivariate multiple linear regression and multivariate non-linear regression are relatively few in research and publications at current stage.

In many cases, there are more than one quantity which needs to be predicted, and a single quantity often corresponds to a set of predictor variables. Under such condition, regressions with prediction equations only relate to each one of the predictor variables seems unreliable. Therefore, it is essential to import the multiple regression to produce a conclusion with higher accuracy. For a simple instance, when it is needed to predict the change in valuations of virtual currency, which is related to over 30

econometric or political variables as predictors. If using a method of single variable regression, the prediction is obviously inaccurate since the variables are strongly corresponded and the fluctuation of the price of each virtual currency will affect each other. Consequently, by taking advantage of the correlations, a more accurate prediction can be produced. Therefore, compared with using only one independent variable, multiple regression is both more practical and effective in predicting dependent variables by the optimal combination of multiple independent variables.

In this article, four kinds of different regressions are discussed: multivariable linear regression, multivariate multiple linear regression, multinomial logistics regression and multivariate non-linear regression. By referring to different articles and examples, it can be found that different regression has their own advantages and disadvantages. Besides, they are also suitable for different situations for their different features.

2. Methods And Applications

2.1. Multivariable/Multiple Linear Regression (MLR)

In a multivariable linear regression model, one variable is usually affected by multiple variables. For example, as for the total expenditure of a family in a year, we not only need to consider the factors affecting disposable income, but also the price level and the interest. So, in this kind of situation, we need to apply multivariable linear regression model.

Now many fields have already applied this model. In manufacturing industry. According to the research of Tsou alas et al., we can use multivariable linear regression and genetic algorithm analysis to predict the occupational risk in shipbuilding industry [7]. Relevant parameters include date and time, individual specialty, incident type, hazardous situation, and risky behavior involved in the incident. A normalization formula is used to estimate the Normalized Risk Index (NRI) of each category concerning the parameters.

$$\mathrm{NRI} = \frac{\sum_i x_i y_i}{3*\sum_i x_i}, i = 1 - 4 \tag{1}$$

This is the percentage of occupational injuries caused in each case, and the corresponding severity. Therefore, the final value of NRI is scaled from 0 to 1. The resulting value of each parameter was ultimately used as the input value of the MVLR-GA model. After replacing the regression coefficient, the final MVLR model equation of occupational risk is shown as follows.

$$f(x) = -0.535 + 0.067\mathrm{DT} + 0.428\mathrm{S} + 0.354\mathrm{I} + 0.602\mathrm{DS} + 0.301\mathrm{DA} \tag{2}$$

Where DT is day and time parameter, S is the specialty, I is the type of incident, DS is the dangerous situation and DA is the dangerous actions involved.

In the humanities and social sciences field, we can apply multiple linear regression to find anxiety, depression and stress among medical undergraduate students and their socio-demographic correlates. The way to do the study is using depression anxiety and stress scale (DASS-21) to estimate the level of anxiety, depression and stress among undergraduate students from different academic institutions in Pokhara Metropolitan [8].

In medical field, we can find the problems of sleep-disordered breathing and insulin resistance in middle-aged and overweight men. In this study, we need to find ways to measure the metabolic consequences and community prevalence of sleep-disordered breathing in mildly obese rather than healthy individuals. The measurements include polysomnography, a multiple sleep latency test, an oral glucose tolerance test, determination of body fat by hydro densitometry, and fasting insulin and lipids. We use multivariable linear regression to find the relation among them according to Punjabi et.al [9]. We can find that multivariable linear regression has a wide application prospect in the future.

2.2. Multivariate Multiple Linear Regression

In order to make good use of correlations between predictive variables while maintaining accuracy even when responses are unrelated, a new technique called the curd and whey method was introduced. In the case of correlation between reactions, its use can significantly reduce the prediction error.

The curds and whey method are a modality of multi-variable shrinkage. It converts matrix, contrasts, and then converts back into inverse matrix. It gains its power by contracting in a right-handed coordinate system and then be seen as a multi-variable summarize of proportional contraction from cross-feasibility [10].

Overall, when there exist n responses $y = (y_1, \ldots, y_n)$ with isolated optimal squares regressions$\hat{y} = (\hat{y}_1, \ldots\ldots, \hat{y}_n)$. The example above suggests the validation that if the corresponding variables are related, we will be able to get more accurate predictions for each response variable by using a linear combination of ordinary least squares predictors with the formula below, rather than using the least squares predictors themselves.

$$\tilde{y}_i = \bar{y}_i + \sum_{k=1}^{n} b_{ik}(\hat{y}_k - \bar{y}_k), i = 1, \ldots, n \tag{3}$$

$$\hat{y}_i = \bar{y}_i + \sum_{j=1}^{n} \hat{a}_{ij}(x_j - \bar{x}_j) \tag{4}$$

The centred proportional contrast $\tilde{y}$ for an isolated responsible variable y can be expressed as Equation (5).

$$\tilde{y} = b\hat{y} = \sum_{j=1}^{n}(b\hat{a}_j)x_j \tag{5}$$

Where $\hat{y}$ and $\{\hat{a}_j\}_1^n$ are the usual least squares estimates. Each ordinary least squares coefficient $\hat{a}_j$ is measured by the same element b, and the general biased estimation is a linear regression of the ordinary least squares answer $\tilde{y}$. Here several formulas have been suggested in order to estimating the extent of contrast to obtain a more accurate mean squared deviation expectation, in which the expected figure lies above the joint distributive function $F(x, y)$ of the predictors x and the respond variable y.

$$E(y - \tilde{y})^2 < E(y - \hat{y})^2 \tag{6}$$

A natural extension of the multivariate adjustment is to express each deviation ordinary least squares estimation as a general linear function. In vector representation, this is represented by an equation in which matrix B is thought of as a compression matrix that converts ordinary least squares estimates to biased estimations. The target is to acquire an estimate matrix B of the best pinched matrix B* whose elements are defined as Equation (7).

$$\{b_{ik}^*\}_{k=1}^{q} = \arg\min(\{\beta_k\}_1^q)[E\{y_i - \sum_{k=1}^{q}\beta_k \hat{y}_k\}^2], i = 1, \ldots, q \tag{7}$$

For an individual respond variable, the usual least square estimation can be better than the bias contraction estimation in the form of accuracy rating. Examples include proportional shrinking, ridge regression, principal component regression, and partial least squares regression. These results suggest that, in the context of the merge shrink process, it is advantageous to treat the set of responses as vector-valued variables.

Curd and whey methods tend to enhance the expected predictive prediction rating of each respond variable. This proposal an interesting prospect that even through there is merely one interesting answer, if there are variables associated with it, the prediction of the interesting answer can be enhanced by importing other variables as extra answers. Certainly, if the figures of these variables can be applied to make predictions, they can also be considered as predictors and covered in the regression equation. However, in several cases, the data is probable to contain measurements of variables that are not valid in the prediction settings.

2.3. Multinomial Logistic Regression

2.3.1. Multiple Logistic Regression

The logarithmic linear model is a combination of the analytical methods of contingency table and linear models, followed by applying the basic idea which is similar to analysis of variance (ANOVA) and the logistic transformation to test the magnitude of the effects of the variables and their interaction effects. The logarithmic linear model is transformed by the logit process to produce the logistic model that analyses the causal relationship between the dependent and independent variables. That is, it can analyze the causal relationship between the measure or category variables and the dichotomous variables.

The logistic regression can be extending to models with multiple explanatory variables. The multiple (binary) logistic regression model has a response variable Y with two measurement levels (dichotomous) and explanatory variable $X(x_1, x_2, \dots, x_n)$, which is the Equation (8).

$$\begin{aligned} \log(\text{odds}) = \text{Logit}(p) = \text{Logit}[P(Y=1)] = \ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = \ln\left(\frac{p}{1-p}\right) \\ = \ln\left[\exp(\beta_0 + \sum_{i=1}^{n}\beta_i x_i)\right] = \beta_0 + \sum_{i=1}^{n}\beta_i x_i \end{aligned} \tag{8}$$

p indicates the probability of the target event happened, while $1-p$ indicates the probability of the opposite event. n represents the number of factors that affect the target event. β_0 means the regression intercept (constant). x_i refers to the i-th influencing factor and β_i shows the regression coefficient of the i-th factor, which can be fitted to the model.

The model for log odds is Equation (9).

$$p = \pi(x) = \frac{\exp(\beta_0 + \sum_{i=1}^{n}\beta_i x_i)}{1+\exp(\beta_0 + \sum_{i=1}^{n}\beta_i x_i)} \tag{9}$$

Multiple logistic regression models have a broad spread of applications. For example, in 2002 [11], Tabaei and Herman modelled a multivariate logistic regression and developed an equation for prediction of undiagnosed diabetes to find the risk factors. In 2014 [12], Mokhtari applied this model based on rock geochemical data to plot the Hydrothermal alteration map. Consequently, this kind of models can be used in statistical analysis in a considerable number of different fields. In fact, polynomial logistic regression models are a fairly simple extension of binary models that rely primarily on logit analysis or logistic regression [13].

2.3.2. Nominal Logistic Regression and Ordinal Logistic Regression

The case of the logistic model discussed earlier for binary data, it is sometimes possible for the response variable to take on three or more values. Depending on the type of response variable, it is divided into nominal logistic regression and ordinal logistic regression [14].

When there are three or more categories with no natural ordering, nominal logistic regression can be used, where the logarithmic probability of the outcome is modeled as a linear combination of predictive variables. When three or more categories have a natural ordering of levels, an alternative formula, called sequential logistic regression, can be used, but the ranking of the levels does not necessarily mean that they are equally spaced.

When the nominal or ordinal corresponding variable has more than one category, the multinomial logit model takes the form of pairing each category with an underlying category, usually taking the final category as the reference.

$$\ln\left(\frac{\pi_j(x_i)}{\pi_J(x_i)}\right) = \ln\left(\frac{p_j}{p_J}\right) = \beta_{j0} + \sum_{i=1}^{n}\beta_{ji}x_i, j = 1,2,3, \dots, J-1 \tag{10}$$

$$p_1 + p_2 + \cdots + p_J = 1 \tag{11}$$

The model for log odds is Equation (12).

$$\log(\pi_j(x_i)) = \frac{\exp(\beta_{j0} + \sum_{i=1}^{n} \beta_{ji} x_i)}{1 + \sum_{j=1}^{J-1} \exp(\beta_{j0} + \sum_{i=1}^{n} \beta_{ji} x_i)} \tag{12}$$

The model parameters are estimated by the method of Maximum Likelihood (ML).

The nominal or ordinal logistic regression method also has various applications when the form of data is appropriate. For example, the paper by Ovaskainen, et al. modelled such multivariate logistic regression of species symbiosis to produce new hypotheses on fungal interactions [15]. The research, conducted by Frangos, et al. revealed the significant risk factors of negative psychological beliefs about the Internet surfing among college students in Greece through building an ordinal logistic regression model [16].

The logistic regression model has two major advantages. Firstly, the logistic regression model can well solve the problem with non-linearity and multiple independent variables and estimate the extent of the influence of the respective variables on the dependent variable. Secondly, the model does not require the variables to obey a multivariate normal distribution and is more widely used.

2.4. Multivariate Non-Linear Regression

To solve the problem with multivariate and non-linear regression, a new method called simplex algorithm is often used, based on ordinary least squares, which is relatively simple, and the convergence effect and convergence speed are ideal. In the case of mastering the ordinary least squares, the key to solving the above problems is to determine the type of curve and how to convert it into a linear model. The determination of curve type is generally considered from two aspects: one is to deduce theoretically or speculate by experience according to professional knowledge, and the other is to determine the general type of curve by drawing and observing scatter diagram when professional knowledge is powerless.

Aiming at estimating the lighting energy demand of rooms with parameter changes in characteristics, the method used in this study is based on a series of statistical analysis are applied to the results of a large number of simulations [17].

To begin with, the parameters affecting the amount of daylight in indoor space and the energy demand of related electric lighting were identified. Then, through dynamic lighting simulation, the annual daylighting conditions, and energy demand of a number of collocations of a target room are analysed. And Daysim, was used for this purpose, which calculates daylight through annual simulations based on dynamic climate.

Table 1. Summary of the design variables

Site	Berlin, Germany (L=52.3°N)		Turin, Italy (L=45.2°N)			Catania, Italy (L=37.5°N)	
Room depth RD[m]	3	4.5	6	7.5	9	10.5	12
Window-to-wall ratio WWE [-]	0.2	0.3	0.4		0.5	0.6	
Obstruction angle Yob [°]	0	15	30	45		60	75
Room orientation OR	south		north			west	
Average target illuminance E [lx]	150		300	500		750	
Lighting control system	on-off manual (MAN)			Automatic daylight responsive (DR)			

The Table 1 gives us a general overview of design variables which is used in Daysim simulations at various stages of the analysis process.

Verso generates a sub database as the basis for statistical analysis to build two mathematical models to predict the energy demand of rooms with manual systems. In statistical analysis, regression techniques are used to define a mathematical model that is suitable for immediately providing a good

estimation of energy demand function. After explaining the explanatory variables such as site and direction, the mathematical model can obtain by non-linear multiple regression analysis. Different nonlinear models are regressed. All regressions are based on the explanatory variables acting on demand for energy after logarithmic behavior. In order to simplify the model into a linear model, the explicative variables and response variables are properly transformed, and then multiple linear regression is performed by SPSS software package.

Statistical analysis including the Mean Bias Error (MBE) (Normalized Mean Bias Error–NMBE), the Mean Squared Error (MSE), the Root Mean Square Error (RMSE) and the coefficient of variation (CV) are also carried out to evaluate how closely the estimated energy demand values is suitable for the simulated values. These statistics are shown as follows:

$$\mathrm{MBE} = \frac{\sum_{i=1}^{828} \mathrm{ED}_{\mathrm{estimated}} - \mathrm{ED}_{\mathrm{stimulated}}}{828} \tag{13}$$

$$\mathrm{NMBE} = \frac{\mathrm{MBE}}{\mathrm{mean}(\mathrm{ED}_{\mathrm{stimulated}})} \tag{14}$$

$$\mathrm{RMSE} = \sqrt{\mathrm{MSE}} = \sqrt{\frac{\sum_{i=1}^{828} (\mathrm{ED}_{\mathrm{estimated}} - \mathrm{ED}_{\mathrm{stimulated}})^2}{828}} \tag{15}$$

$$\mathrm{CV} = \frac{\mathrm{RMSE}}{\mathrm{mean}(\mathrm{ED}_{\mathrm{stimulated}})} \tag{16}$$

However, Pellegrino et.al. used multivariate non-linear regression to solve the problem successfully. In many cases without the powerful professional knowledge, it is hard to get the right model.

3. Conclusion

In this paper, the equations and applications of four types of different regressions are discussed. Each kind of regression is suitable for different situations by searching for examples and articles. Multivariable linear regression is widely used in many fields such as manufacturing industry, the humanities and social sciences field and medical field. In the articles quoted, the phenomenon affected by multiple variables can be modelled as a multivariable linear regression. As for multivariate multiple linear regression, more than one quantity is required to be predicted and a quantity always depends on multiple predictor variables to provide a higher accuracy. When it comes to multinomial logistics regression, several regression models of deformation which changes with different data types are mentioned. The transformed model has a wide application range. The last one, multivariate non-linear regression, needs a variety of professional knowledge to model. To conclude, the multiple regression is an essential approach when predicting response variables with more than one explanatory variable.

References

[1] A. Bakst, A modification of the computation of the multiple correlation and regression coefficients by the Tolley and Ezekiel method, in: Journal of Educational Psychology - J EDUC PSYCHOL, vol. 22, ResearchGate, 1931, pp. 629 – 35. DOI: https://doi.org/10.1037/h0075404.

[2] M. S. Bartlett, Further aspects of the theory of multiple regression, in: Mathematical Proceedings of the Cambridge Philosophical Society, vol. 34, 1938, pp. 33 – 40. DOI: https://doi.org/10.1017/S0305004100019897.

[3] J. D. Elashoff, F. N. Kerlinger, E. J. Pedhazur, Multiple Regression in Behavioral Research, in: Journal of the American Statistical Association, vol. 70, 1975, p. 959. DOI: https://doi.org/10.2307/2285468.

[4] R. Mac Nally, Multiple regression and inference in ecology and conservation biology: further comments on identifying important predictor variables, in: Biodiversity & Conservation, vol. 11, Springer Link, 2002, pp. 1397 – 401. DOI: https://doi.org/10.1023/A:1016250716679.

[5] A. E. Maxwell, Limitations on the use of the multiple linear regression model, in: British Journal of Mathematical and Statistical Psychology, vol. 28, 1975, pp. 51 – 62. DOI: https://doi.org/10.1111/j.2044-8317.1975.tb00547.x.

[6] D. W. Hosmer, S. Lemesbow, Goodness of fit tests for the multiple logistic regression model, in: Communications in Statistics - Theory and Methods, vol. 9, 1980, pp. 1043 – 69. DOI: https://doi.org/10.1080/03610928008827941.

[7] V. D. Tsou alas, N. G. Fragiadakis, Prediction of Occupational Risk in the Shipbuilding Industry Using Multivariable Linear Regression and Genetic Algorithm Analysis, in: Safety Science, vol. 83, ScienceDirect, 2016, pp. 12 – 22. DOI: https://doi.org/10.1016/j.ssci.2015.11.010.

[8] S. Iqbal, S. Gupta, E. Venkatarao, Stress, anxiety & depression among medical undergraduate students & their socio-demographic correlates, in: The Indian Journal of Medical Research, vol. 141, 2015, pp. 354 – 57.

[9] N. M. Punjabi, J. D. Sorkin, L. I. Katzel, A. P. Goldberg, A. R. Schwartz, P. L. Smith, Sleep-disordered Breathing and Insulin Resistance in Middle-aged and Overweight Men, in: American Journal of Respiratory and Critical Care Medicine, vol. 165, 2002, pp. 677 – 82. DOI: https://doi.org/10.1164/ajrccm.165.5.2104087.

[10] L. Breiman, J. H. Friedman, Predicting Multivariate Responses in Multiple Linear Regression, in: Journal of the Royal Statistical Society: Series B (Statistical Methodology), vol. 59, Wiley Online Library, 1997, pp. 3–54. https://doi.org/10.1111/1467 - 9868.00054.

[11] B. P. Tabaei, W. H. Herman, A Multivariate Logistic Regression Equation to Screen for Diabetes, in: DIABETES CARE, vol. 25, 2002, p. 5.

[12] A. R. Mokhtari, Hydrothermal alteration mapping through multivariate logistic regression analysis of lithogeochemical data, in: Journal of Geochemical Exploration, vol. 145, ScienceDirect, 2014, pp. 207–12. DOI: https://doi.org/10.1016/j.gexplo.2014.06.008.

[13] A. M. El-Habil, An Application on Multinomial Logistic Regression Model, in: Pakistan Journal of Statistics and Operation Research, vol. 8, 2012, pp. 271 – 91. DOI: https://doi.org/10.18187/pjsor.v8i2.234.

[14] D. Hedeker, Multilevel Models for Ordinal and Nominal Variables, in: Handbook of Multilevel Analysis, edited by Jan de Leeuw and Erik Meijer, Springer, 2008, pp. 237 – 74. DOI: https://doi.org/10.1007/978-0-387-73186-5_6.

[15] O. Ovaskainen, J. Hottola, J. Siitonen, Modeling species co-occurrence by multivariate logistic regression generates new hypotheses on fungal interactions, in: Ecology, vol. 91, Wiley Online Library, 2010, pp. 2514 – 21. DOI: https://doi.org/10.1890/10 - 0173.1.

[16] C. C. Frangos, C. C. Frangos, L. Sotiropoulos, Problematic Internet Use among Greek University Students: An Ordinal Logistic Regression with Risk Factors of Negative Psychological Beliefs, Pornographic Sites, and Online Games, in: Cyberpsychology Behavior & Social Networking, vol. 14, Baidu Scholar, 2011, p. 51. DOI: https://doi.org/10.1089/cyber.2009.0306.

[17] V. R. M. Lo Verso, A. Pellegrino, and F. Pellerey, A multivariate non-linear regression model to predict the energy demand for lighting in rooms with different architectural features and lighting control systems, in: Energy and Buildings, vol. 76, 2014, pp. 151 – 63. DOI: https://doi.org/10.1016/j.enbuild.2014.02.063.