Journal of The West Bengal University of Health Sciences October 2020 Vol 1 Issue 2

Original Article

Understanding the application of multiple linear regression model in health care research using simulation with computer generated data.

Keshab Mukhopadhyay[1], Ritesh Singh[2], Chanchal Kumar Dalai[1], Shah Newaz Ahmed[1]✉, Kushal Banerjee[3]

ABSTRACT

Background: The exponential increase in medical information in contemporary science warrants the use of computational tools to simplify and ameliorate patient care. Multiple linear regression modelling is a statistical method that has wide applications in analysis and interpretation of clinical data. In this article, we describe the method of developing a multiple linear regression model using simulation of computer generated data.

Methods: Data was generated for a sample size of 40, for one dependent variable (Y) and four independent variables (X1, X2, X3 and X4). In the first step, bivariate correlation was used to find the individual strength of correlation (R) between the dependent and independent variables. In the second step, the significant variables were added in the model in order of decreasing value of R. Variables which remained statistically significant ($p<0.1$) in the model were retained while insignificant and multicollinear variables were removed. The final best fit model was conceived with the significant predictors.

Results: The R value for variables X1 to X4 was 0.933, 0.911, 0.725 and 0.148 respectively. X1, X2 and X3 were statistically significant ($p<0.001$) while X4 was non-significant ($p=0.36$). X2 and X4 were not included in the best fit model because of multicollinearity and statistical non-significance respectively. The best fit model was represented by the equation

1. Dept. of Pharmacology, College of Medicine and JNM Hospital, The West Bengal University of Health Sciences, Kalyani, Nadia, West Bengal
2. Dept. of Community and Family Medicine, All India Institute of Medical Sciences, Kalyani, Nadia, West Bengal.
3. Dept. of Pharmacology, Medical College and Hospital, Kolkata, West Bengal

✉ email: shahnewazpharmacology@gmail.com

Received: 21 September 2020

Accepted: 11 October 2020

Published online: 16 October 2020

Citation: Mukhopadhyay K, Singh R, Dalai CK, Ahmed SN, Banerjee K. Understanding the application of multiple linear regression model in health care research using simulation with computer generated data. J West Bengal Univ Health Sci. 2020; 1(2):44-52.

Y=0.692*X1+0.218*X3+2.003 where 0.692 and 0.218 were the unstandardized coefficients for X1 and X2 respectively and 2.003 was the constant.

Conclusion: Multiple linear regression modelling can be a useful tool for studying the simultaneous effect of multiple variables on a single dependent variable.

Key words: mathematical modelling, multiple linear regression, simulation, statistical method

Introduction:

In research work, a mathematical method known as simple linear regression is of great utility when a strong correlation exists between a dependent variable and a single independent variable. This allows the researcher to accurately predict the value of the response variable with the independent variable when the value of the former is not available.However, in the world of biological science, heterogeneity of data is the rule rather than exception. This reflects asgaps in explanation of a clinical entity when attempts are made to explain the same with a single variable.[1] This warrants the collating of multiple independent variables into a single statistical model to predict or explain the dependent variable. Such a method known as multiple linear regression is a very useful tool in medicine to perform a variety of functions ranging from assessment of risk factors to arriving at a diagnosis.[2] The method begins with identification of the significant predictors by separately assessing the strength of correlation between the predictors and the dependent variable. In the second step, the significant correlates are tested for inclusion in a collective model using proper statistical techniques. The model is deemed "best fit" when it explains the maximum proportion of the variability for the dependent variable with the minimum statistical error. In this article, we describe a step by step procedure to construct a multiple linear regression model using computer generated data for multiple variables.

Methods:

Generation of Data

A set (n=40) of data was generated for five variables(Y, X1, X2, X3 and X4) in Microsoft Excel using the function I=RANDBETWEEN(A,B) where I is a random integer between A and B (Table 1). The set of 40 values for the variables Y, X1 and X2 was generated in 4 incremental blocks with random integers ranging between 1-10, 11-20, 21-30 and 31-40 for the first, second, third and fourth block respectively. Similarly, the data for the variable X3 was generated in 2 blocks ranging between 1-20 and 21-40 for the first and second block respectively, and the data for X4 was generated in a single block ranging between 1-40.

Correlation Statistics

The dependent variable was Y and the independent variables were X1, X2, X3 and X4. Bivariate correlation was employed to test the strength of correlation between the dependent and independent variables. The strength of correlation was obtained separately for each of the independent variables.The strength of correlation was assessed using Pearson's correlation coefficient (r). A p value of <0.05 was deemed statistically significant.

Simulation of Multiple Linear Regression Model

All simulation work was done in SPSS version 22.0. The construction of the model was initiated with a simple linear regression

Table 1: Computer generated data for the dependent (Y) and the independent variables (X1, X2, X3 and X4)

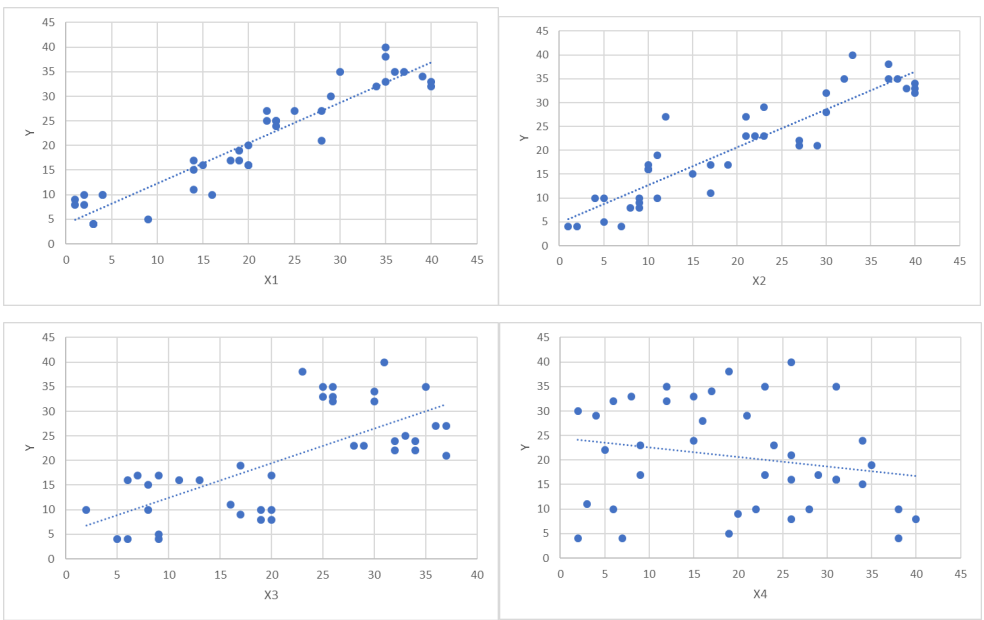
Sl. No.	Y	X1	X2	X3	X4
1	4	3	1	5	38
2	10	4	9	20	38
3	5	9	5	9	19
4	10	2	5	19	6
5	10	4	4	8	28
6	4	3	7	9	7
7	8	2	9	20	26
8	8	1	8	19	40
9	4	3	2	6	2
10	9	1	9	17	20
11	15	14	15	8	34
12	17	14	17	7	9
13	17	18	10	9	29
14	16	20	10	13	31
15	10	16	11	2	22
16	17	19	19	20	23
17	16	20	10	11	26
18	19	19	11	17	35
19	16	15	10	6	31
20	11	14	17	16	3
21	24	28	12	29	2
22	20	23	23	32	15
23	22	22	22	33	21
24	21	28	27	37	4
25	27	29	21	36	26
26	27	22	23	34	34
27	27	23	30	28	5
28	22	20	29	34	24
29	27	23	21	32	9
30	25	25	27	37	16
31	32	40	30	30	6
32	35	37	32	35	23

Sl. No.	Y	X1	X2	X3	X4
33	40	35	33	31	26
34	32	34	40	26	12
35	33	40	40	25	8
36	35	30	38	25	12
37	33	35	39	26	15
38	38	35	37	23	19
39	34	39	40	30	17
40	35	36	37	26	31

Table 2: Stepwise simulation of multiple linear regression for construction of the best fit model.

Step	Variables added	P value	R^2	Adjusted R^2	Inter-variable R	Variable removed
01	X1	<0.001	0.871	0.868		
02	X1 X2	<0.001 0.001	0.903	0.898	0.89	X2
03	X1 X3	<0.001 0.003	0.898	0.893	0.64	

Figure 1: Scatter plot between the dependent variable (Y) and the independent variables (X1, X2, X3 and X4)



model between the strongest predictor and the dependent variable. The other significant correlates were added sequentially into the model in decreasing order of the strength of correlation. The model was checked for increase in explained variability (Coefficient of determination, R^2) and change in significance level of the independent variables at each step of addition of a new variable. The threshold probability below which the variable was retained in the model was set at 0.1. Multicollinearity was checked by inter-correlating the independent variables. A r valueof greater than 0.7 between the independent variables was set as the test limit of multicollinearity and the weaker correlate was droppedfrom the model.The model was checked for normality of data and homogeneity of variances using Kolmogorov-Smirnov test. A p value greater than 0.05 was deemed to be indicative of normality.

In the end, the best fit model was constructed with the retained significant predictors.

Results:

The dependent variable Y was normally distributed. The r value for bivariate correlation between Y and X1, X2, X3 and X4 was 0.933, 0.911, 0.725 and -0.148 respectively (Figure 1). While the strength of correlation for X1, X2 and X3 was statistically significant ($p<0.001$), it was statistically non-significant for X4 ($p=0.36$). The most significant correlate was X1 (R=0.933 and 0.911 for X1 and X2 respectively). The construction of multiple linear regression models began with a simple linear regression model between Y and X1. The simple linear equation for X1 was Y = 0.812*X1+4.226. X1, X2 and X3 were added sequentially to the model in order of strength of correlation. X4 was not added in the model as it was not a significant correlate. The change in adjusted R^2 and the significance level after each addition is shown in Table 2. X2 was found to be a multi-collinear variable (r with X1=0.89) and dropped from the model. The residuals were seen to obey the normal distribution pattern of data.The best fit model was developed with the significant correlates after the entry and elimination steps and was represented with the equation Y=0.692*X1+0.218*X3+2.003 where 0.692 and 0.218 were the unstandardized coefficients for X1 and X2 respectively and 2.003 was the constant. The proportional of explained variability by the best fit model was 89.8 %.

Discussion:

The Prelude

The history of medicine dates back to time immemorial. The birth of the science took place the very day the prehistoric man nurtured the zeal to care, soothe and alleviate the pain of a fellow human being suffering from disease or infirmity. Since then,medical science has witnessed the transition from a primitive intuitive science to the present day modern medicine where diagnosis and treatment of disease is evidence based.[3] However, the realms of variability in human traits and environmental exposure poses a serious challenge to diagnostic accuracy and therapeutic homogeneity.[4] Traditionally, both diagnosis and response to therapydepends on a set of clinical, laboratory and genetic determinantswhere we see a convergence of multiple factors when reaching a diagnosis and divergence of treatment efficacy after therapeutic initiation (Fig 2). The boom in medical knowledge in contemporary times has transformed patient care into an exhaustive and mindboggling data-driven task which requires collating of information from multiple sources.[5] Heuristically, we envision an unavoidable need to empower ourselves with computational and statistical knowhow to acclimatise ourselves to the information era in the best interests of mankind.

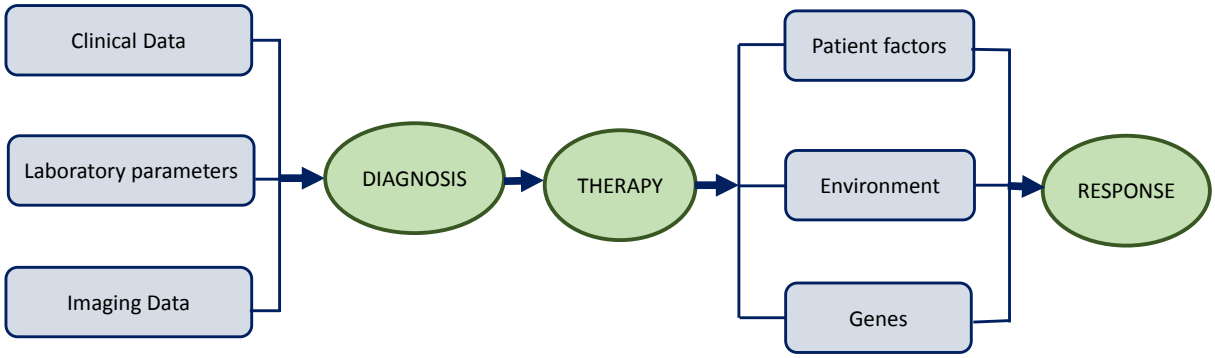


Figure 2: Flow diagram showing convergence of information leading to diagnosis and divergence of response to therapy after therapeutic initiation.

The Method

In this article, we have described the statistical method of multiple linear regression modelling which is an indispensable tool in performing and solving an array of tasks in the field of medicine. The method has the potential to define and attain diagnostic and therapeutic objectives with a maximised non-redundant mathematical accuracy. An optimum number of independent variables are incorporated simultaneously into a predictive model to obtain the maximum proportion of explained variability for the outcome variable. The contribution of the individual predictors is pooled into a single common final effect.[6] Though it is also possible to introduce categorical or ordinal variables into the model as predictors, for simplicity of understanding, we have used only continuous independent variables in the simulation.

The Model

In our model, the proportion of explained variability (PEV) for the dependent variable Y was 87.1% with the strongest correlate X1. The PEV increased to 89.8% when X3 was added to the model. X2 was also a very strong correlate to the dependent variable but as per pre-set criteria for multicollinearity was not retained in the final best fit model. X4 was not included in the model because it was not a statistically significant correlate as per the cut-off probability level.

Multicollinearity Diagnostics

Multicollinearity is a collateral statistical load in multiple linear regression modelling due to high level of inter-correlation between the predictors that does not contribute significantly to the PEV of the model but increases the margin of error of the parameter estimates of the regression coefficients, thereby attenuating the external validity of the model. Detecting and eliminating multicollinearity is an important exercise in multiple linear regression and is achievable by multiple established statistical techniques. In our simulation work, we used a simple and easy way to detect and eliminate the multi-collinear variables. Barring the most significant correlate with which the model was incepted, a predictor which has a Pearson's correlation coefficient greater than a cut-off value (pre-set at 0.7 in our model) with any of the other predictors was deemed a multicollinear variable and dropped from the model. The other popular techniques of diagnosing multicollinearity includes variance inflation factor (VIF), tolerance, condition indexing, condition numbering, Farrar-Glauber test,

variance decomposition proportion and others.[7] Tolerance and VIF of a particular predictor for inclusion in a model depend on the PEV (Rp^2) of the predictor (which is now the dependent variable) by the remaining predictors included in the model. The value of tolerance is calculated as 1- Rp^2. VIF is the inverse of tolerance. The cut-off value for tolerance and VIF for a predictor to be included in the model is conventionally set at or above 0.2 and 5 respectively.[7] The discussion of multicollinear diagnostics has a much bigger scope, which is judiciously kept outside the purview of this article.

Model Flexibility

The pre-specified values of multicollinearity between predictors, significance level and increase in PEV on feeding of predictor into model are non-stringent statistical criteria which are best judged by researchers to maximise the clinical utility of the model without violating the basic principles.The aim of the best fit model is to maximise the magnitude of PEV with minimum redundancy and the least proportion of error so that the model subserves internal as well as external validity.

Simulation in Biological Research

Our work not only aims to describe multiple linear regression in a simple methodical way but also seeks to familiarise a basic health care researcher with the concept of simulation. Simulation before site work can predict the viability and feasibility of the proposed research work and can save a lot of time and money. Our work shows that simulation can be done with readily available tools like MS Excel and SPSS, even without the knowledge of coding in High Level Programming Languages. In a first of its kind study, Balakrishnan et al, simulated a model for standard quantification of drug consumption in paediatric population using computer generated data in Microsoft Excel. The favourable output of the simulation led the researchers to real world validation of the method in a pilot cohort. The pioneering research paved the way for development of a standard unit of drug consumption in paediatric population which ensures comparability and uniformity of drug consumption quantification across heterogenous paediatric samples.[8] Simulation has also been used in the arena of comparative cost-effective analysis. Kongnakorn et al, using public domain inputs for infection rate, resistance pattern, recovery rate, adverse event and cost of therapy simulated a comparative analysis between three different treatment armsin complicated abdominal infectionsmimicking a 5-year timeline of 5000 virtual patients. The premonitory results of the computer model not only guide intensivists in the choice of therapy butcan also be a source of immense utility for policy makers in resource allocation.[9] The aforementioned two exampleson computer simulation, of the myriad available in literature are beckoning evidence of the existing usefulness and future potential of simulation in healthcare research.

Brief Literature Review

Multiple linear regression is not new in medical research. Ahmed et al, developed a multiple linear regression model for prediction of left ventricular mass (LVM) of the heart using a set of electrocardiographic (ECG) and clinical variables. The prediction of LVM with ECG only as a point-of-care tool, without any resource-dependent imaging modality, can serve as a ready and cost-effective alternative in resource-crunched set-ups and can ease financial and temporal burden.[6] Multiple linear regression can also be used to ascertain the risk factors of disease and morbidity. In a countrywide database analysis of invasive meningococcal infection, the researchers observed a strong correlation between the incidence of the infection and the carrier population and

the percentage of susceptible in the sub-population.[10] Risk factors as well severity of both communicable and non-communicable disease have been determined using the statistical method of multiple linear regression modelling in numerous research work. The method also helps in finding out the determinants of response to therapy in a particular health condition, provided the outcome (response to therapy) is a continuous variable. In a study, seeking to judge the efficacy of dulaglutide as on add-on therapy to insulin in uncontrolled type 2 diabetes mellitus, the researchers tested multiple parameters for a possible association with the magnitude of glycosylated haemoglobin at the end of 6 months of treatment. Baseline Hba1c was found to be a significant correlate of Hba1c reduction in a fixed effect multiple linear regression model.[11] The list of examples of multiple linear regression in medical literature is exhaustive and its ramifications have ubiquitous applications in medical research.

Limitations

The number of predictor variables (04) and the sample size (n=40) in the simulation was kept on the lower side to enhance the ease of understanding and minimise mathematical conundrum. In real-world settings, the data is invariably more complex. A priori calculation of sample size, using the optimal power, limit of type I error and the anticipated effect size should be undertaken. The selection of optimum number of explanatory variables commensurate with sample size is also essential to avoid overfitting or underfitting of the model.In addition, categorical and ordinal data can also be introduced as explanatory variables in a multiple linear regression model by recoding into discrete numerals. However, our work was confined to continuous predictors only. Finally, identification of confounding variables is important in any biological research to avoid equating correlation to causation. All mathematical relations must be judged in conjunction with the biological plausibility.

Conclusion:

Multiple linear regression modelling is a useful method for assessment of multiple clinical end-points in modern medicine and can be effectively employed for development of ready-made tools for risk assessment, establishing diagnosis or monitoring therapeutic efficacy in patient care.

References:

1. Slinker BK, Glantz SA. Multiple linear regression: accounting for multiple simultaneous determinants of a continuous dependent variable. Circulation. 2008; 117(13):1732–7.
2. Eberly LE. Multiple linear regression. Methods Mol Biol. 2007; 404:165–87.
3. Djulbegovic B, Guyatt GH. Progress in evidence-based medicine: a quarter century on. Lancet. 2017; 390(10092):415–23.
4. Cahan A, Cimino JJ. A learning health care system using computer-aided diagnosis. J Med Internet Res. 2017; 19(3):e54.
5. Yu K-H, Beam AL, Kohane IS. Artificial intelligence in healthcare. Nat Biomed Eng. 2018; 2(10):719–31.
6. Ahmed SN, Jhaj R, Sadasivam B, Joshi R. Prediction of left ventricular mass index using electrocardiography in essential hypertension - a multiple linear regression model. Med Devices (Auckl). 2020; 13:163–72.
7. Kim JH. Multicollinearity and misleading statistical results. Korean J Anesthesiol. 2019; 72(6):558–69.
8. Sadasivam B, Malik S, Atal S, Ahmed SN. Development and validation of

a mathematical model to quantify antibiotic consumption in paediatric population: A hospital-based pilot study. J Clin Pharm Ther. 2020. DOI: 10.1111/jcpt.13216 [online ahead of print]

9. Kongnakorn T, Eckmann C, Bassetti M et al. Cost-effectiveness analysis comparing ceftazidime/avibactam (CAZ-AVI) as empirical treatment comparing to ceftolozane/tazobactam and to meropenem for complicated intra-abdominal infection (cIAI). Antimicrob Resist Infect Control. 2019; 8:204.
10. Mokhort H. Multiple linear regression model of meningococcal disease in Ukraine: 1992-2015. Comput Math Methods Med. 2020. DOI: 10.1155/2020/5105120.
11. Lee J, Cho YK, Kim HS, Jung CH, Park J-Y, Lee WJ. Dulaglutide as an add-on to insulin in type 2 diabetes; clinical efficacy and parameters affecting the response in real-world practice. Diabetes Metab Syndr Obes. 2019; 12:2745–53.