Journal of Applied Dental and Medical Sciences
NLM ID: 101671413 ISSN:2454-2288
Volume 2 Issue 1 January - March 2016

Review Article

Regression Analysis – Its Formulation and Execution In Dentistry

Sandhya Jain [1], Sunny Chourse [2], Soumya Dubey [3], Sandesh Jain[4], Juhi Kamakoty [5], Deshraj Jain [6]

[1]*Department of Orthodontics, Government Dental College, Indore*
[2]*Post Graduate Student, Dept. of Orthodontics, Government Dental College, Indore*
[3,4] *Intern , Sri Aurobindo College of Dentistry , Indore*
[5] *Acropolis Faculty of Management and Research, Indore*
[6] *Department of Prosthodontics, Government Dental College, Indore*

ARTICLE INFO

Keywords:
Regression Analysis, Linear Regression, Logistic Regression Statistics, Research Methodology

ABSTRACT

Prediction and estimation is the mainstay in the treatment planning in dentistry. With variations being common is many events of the oral cavity, it becomes important to have a methodology which can help us predict the happenings of the region in relation to each other. Regression analysis is one such concept which explores the relationship between two or more quantifiable variables so that one variable can be predicted from other. The aim of this article is to provide a simple yet holistic approach to the understanding of the concepts of Regression Analysis along with its use and misuse, advantages and disadvantages pertaining to the art and science of dentistry.

Introduction

Regression analysis is a tool in assessing specific forms of relationship between the variables. The ultimate objective of this method of analysis is to predict or estimate the value of one variable corresponding to a given value of another variable. The ideas of regression were first elucidated by the English scientist Sir Francis Galton (1822–1911) in reports of his research on heredity—first in sweet peas and later in human stature.[1] He described a tendency of adult offspring, having either short or tall parents, to revert back toward the average height of the general population. He first used the word reversion, and later regression, to refer to this phenomenon. Correlation which is often wrongly confused with regression is concerned with measuring the strength or degree of the relationship between variables. The variables involved in regression can be of either continuous or discrete type. If they are discrete then they should be treated as continuous. Only quantitative variables can be used in regression analysis. It can't be applied to qualitative variables.

THE SCATTER PLOT

The scatter plots aid us in determining the nature of the relationship and correlation by comparing two sets of data. Its gives us a visual picture of any connection between the two variables. Figure I explain the degree and types of correlations derived from the scatter diagram.

* *Corresponding author:* ***Dr Sandhya Jain, MDS, Professor and Head, Dept. Of Orthodontics and Dentofacial Orthopedics Government Dental College, Indore (M.P.) 452001 Mobile No - +91-9425045455 Email: researchorthodontics@gmail.com***

Degree of correlation

None – this is no relationship between the variables.
Low – there is some relationship between the variables but a weak one.
High – there exists a very close relationship between the two variables.
Perfect – it's an ideal relationship. As scores on one of the two variables increase or decrease, the scores on the other variable increase or decrease by the same magnitude

Types of correlation

Positive – the two variables changes in the same direction and in the same proportion.
Negative – the two variables changes in the opposite direction and in the same proportion.
Curved – it represents non linear association between the two variables. Even if the relationship is strong the correlation coefficient can be small or zero.
Partial – it measures the association between two variables while controlling or adjusting the effect of one or more variables.

When a scatter plot indicates there is a high positive relationship between the two variables, it is confirmed by calculating the correlation coefficient, which is deemed to come high. Then only the regression analysis is performed.

COMPONENTS OF REGRESSION ANALYSIS [2,3]

To understand the regression equation, it is first important to know about the various components involved in performing a regression analysis. In a simple regression problem, we are primarily interested in a statistical relationship between one **dependent variable Y** and one **independent variable X**. Dependent variable Y is also called response variable or outcome variable. Independent variable X is called predictor variable, regressor variable, or explanatory variable. The independent variable predicts the value of dependent variable for a given value of independent variable.

The general equation for a straight line may be written as

$$Y = A + B(X)$$

Where,
Y is a value on the vertical axis (dependent variable),
X is a value on the horizontal axis (independent variable),
A is the point where the line crosses the vertical axis and the value of Y at X = 0,
Y-intercept is the value of A when X = 0,
B shows the amount, by which Y changes for each unit change in X, i.e. slope of the straight line.
Figure II explains these variables in a graphical format.

THE FORMULATION OF REGRESSION EQUATION

Collection of data becomes important in research methodology. Data can be qualitative or quantitative. The qualitative study aims to explore and obtain insight, into complex issues such as reason for people attitude or behavior. The results are described in words other than number. Quantitative studies aim to test a hypothesis. The results are given in number of proportion. Quantitative study design may be used to describe how often the event occurs.[4] Statistically significant does not necessarily mean clinically important. It is the size of effect that determines

clinical importance and not the presence of statistical significance.[5]

The purpose of a regression analysis is to find a reasonable regression equation to predict the average value of a response/dependent variable Y that is associated with a fixed value of one independent variable X. If more than one independent variable were used to predict the average value of a response variable, then we would need multiple regressions. Now when we replace the general equation of a straight line with a regression equation we get –

$Y_i = \beta_0 + \beta_1 X_i$

Where,

Y_i is the value of the dependent variable at the *i* th level of the independent variable,

β_0 and β_1 are unknown regression coefficients whose values are to be estimated,

X_i is a known constant, which is the value of the independent variable at the *i* th level.

Let us try to understand this concept through a detailed example. Suppose we want to predict the Lower Anterior Facial Height (LAFH) from the Mandibular Plane Angle (MPA). The MPA is the primary factor i.e. the independent variable (X_i) on the basis of which we want to predict LAFH which is considered to be the dependent variable (Y_i). In 10 subjects the degree of MPA and LAFH was observed and noted down in Table I.

We enter the following data –

X – 17 21 24 20 27 22 23 19 25 28

Y – 56 61 65 60 70 58 63 58 67 72

Now, using the various statistical softwares like SPSS, Minitab, MaxStat, Analytica the values of β_0, β_1 regression coefficients and equation are calculated.

The graph of the regression line is represented in figure III-

The equation of the regression line is –

$Y_i = 29.632 + 1.476 (X_i)$

$Y_i = \beta_0 + \beta_1 X_i$

Here, β_0 is 29.632 which is the value of Y when X is 0. But in this case the MPA can never be practically zero. β_1 is 1.476 which is the amount of changes in Y for each one unit change in X. So, by this regression equation we can conclude that 1 degree increase in Mandibular Plane Angle increases the Lower Anterior Facial height by 1.4764 mm. Also we can calculate the predictable value of LAFH from a given MPA by the equation.

E.g. – What should be the LAFH at MPA of 30 degree?

According to the equation,

$Y_i = 29.63 + 1.4764 X_i$

$= 29.63 + 1.4764 \times 30$

$= 73.922$ mm

So, at MPA 30 degrees the expected LAFH should be 74 mm.

This way the values of the dependent variables can be calculated.

TYPES OF REGRESSION ANALYSIS

1. Bivariate regression – it is the simplest form involving the prediction of the value of unknown variable from the value of known variable. E.g. – to predict the amount of mandibular growth remaining from the Cervical Vertebra Maturation stage.

2. Multivariate regression – as the name suggests, it involves the multiple variables in the regression model. It is used to learn about the relationship between several independent or predictor variables and a dependent or outcome variable. E.g. – when predicting the changes in mandibular length (dependent variable) using IGF-1, cervical stage, skeletal classification, and gender (independent variables).

TYPES OF REGRESSION MODEL

1. Simple Linear Regression Model –

It is the most widely known model in which the dependent variable is continuous and the independent variable can be continuous or discrete. Here the nature of regression line is linear which establishes a relationship between the variables using a best fit straight line, known as the regression line.

In a simple linear regression model we assume that the graph of the mean of the response variable Y_i for given values of the independent variable X_i is a straight line.

2. Multiple Linear Regression Model –

The difference between simple linear regression and multiple linear regression is that the later has more than one (>1) independent variables, whereas the former has only 1 independent variable. Table II explains the simple and multiple linear regression model.

LOGISTIC REGRESSION

It is used when the dependent variable is binary/nominal or categorical with just two values like yes/no, true/false, male/female, healthy/diseased etc. The logit function is a link which provides the distribution amongst these two values ranging from 0 to 1. There are two types of logistic regression. Simple logistic regression with only one independent variable and multiple logistic regressions which has more than one independent variable. Simple logistic regression finds the equation that best predicts the value of the *Y* variable for each value of the *X* variable. What makes logistic regression different from linear regression is that you do not measure the *Y* variable directly; it is instead the probability of obtaining a particular value of a nominal variable. Table III explains the simple and multiple logistic regressions.

DIFFERENCE BETWEEN CORRELATION AND REGRESSION

Correlation measures the strength and direction of relationship between two variables. A parametric correlation coefficient (*r*) is a measure of the linear relationship between two continuous variables. The range of *r* is 1 to –1. When *r* is equal to zero, there is no relationship between the two variables. Regression coefficients estimate how much of the change in the outcome is associated with changes in the explanatory variables. So in simpler words it can be said that, correlation measures the strength of relationship and regression measures the magnitude of relationship between the two variables.

SIGNIFICANCE IN DENTISTRY [6]

Here is Table IV illustrating some examples which explains in respect to orthodontics that how independent variable predicts the value of dependent variable.

ADAVNTAGES OF REGRESSION ANALYSIS

1. Predicts the future course of events in growth, development and treatment.
2. Support treatment planning decisions.

3. Correcting errors.
4. Provide new insights.

LIMITATIONS OF REGRESSION ANALYSIS

1. Looks at only linear straight line relationship between the dependent and independent variable.
2. Observes on the mean of the dependent variable.
3. Sensitive to extreme data.
4. Data must be independent.
5. Not useful when the actual and exact mathematical relationship between the

variables are know. For e.g. Simple algebra shows that change in ANB are

equal to changes in SNA minus changes in SNB. Therefore, it is unnecessary to use multiple regression to estimate the effects of the changes in SNA and SNB on the change in ANB, because we actually know the exact mathematical (linear) relationship among the 3 variables [17].

CONCLUSION

In the formulation and execution of an dental treatment plan, the variables involved in the decision making are often poorly characterized and incompletely validated. For these reasons we have to rely on the mean values or go for a wild guess. The regression analysis is a statistical technique that deals with the analysis of relationship between such variables. They help us to predict the duration, course and outcome of the treatment parameters. It is a complex process of predicting/estimating the magnitude of some unknown characteristics which might be involved in the growth and treatment of a given patient.

REFERENCES

1. Robert R. Sokal and F. James Rohlf. Introduction to Biostatistics. 2nd Ed. Dover publications; 2009.
2. Wayne W. Daniel. Biostatistics A foundation for analysis in the health sciences. 9th ed. Wiley Press; 2009.
3. Chap T. Le. Introduction to Biostatistics. 1st ed. Wiley-Interscience Publications; 2003.
4. Jain S, Sharma N, Jain D. Basic Fundaments of Designing A Quality Research. J Adv Med Dent Scie Res 2015;3(1):88-95.
5. Jain S, Ashaiya A, Chourse S, Jain D. An Overview of Research Methodology Pertaining to Prosthodontics. Ann. Int. Med. Den. Res. 2016;2(1):9-14.
6. Lysle E. Johnston. Regression: Is Your Guess as Good as Mine? Semin Orthod 2002;8:87-91.
7. Burhan and Nawaya: Prediction of unerupted canines and premolars in a Syrian sample. Progress in Orthodontics 2014 15:4.
8. Linge B O, Linge L. Evaluation of the risk of root resorption during orthodontic treatment: a study of upper incisors. Eur J Orthod 1983; 5: 173–183.
9. Wees, Julie Marie. "Short lower anterior face height: phenotypic diversity." MS (Master of Science) thesis, University of Iowa, 2015.
10. Masoud et al. Predicting changes in mandibular length and total anterior facial height using IGF-1, cervical stage, skeletal classification, and gender. Progress in Orthodontics (2015) 16:7.
11. Mack et al. Relationship between body mass index percentile and skeletal maturation and

dental development in orthodontic patients. Am J Orthod Dentofacial Orthop 2013;143:228-34.

12. Crawford TP. A multiple regression analysis of patient cooperation during orthodontic treatment. American Journal of Orthodontics. 1974 Apr 30;65(4):436-7.
13. Nirunsittirat A, Pitiphat W, McKinney CM, DeRouen TA, Chansamak N, Angwaravong O,Patcharanuchat P, Pimpak T. Adverse birth outcomes and childhood caries: a cohort study.Community Dent Oral Epidemiol 2016.
14. Čanković M, Bokor-Bratić M, Novović Z. Stressful Life Events and Personality Traits in Patients with Oral Lichen Planus. Acta Dermatovenerol Croat. 2015 Dec;23(4):270-6.
15. van der Tas JT et al. Association between Bone Mass and Dental Hypomineralization. J Dent Res. 2016 Jan 8. (In Press).
16. Hedström L, Albrektsson M, Bergh H. Is there a connection between sublingual varices and hypertension? BMC Oral Health. 2015;15:78.
17. Tu, Nelson-Moon, and Gilthorpe. Misuses of correlation and regression analyses in orthodontic research: The problem of mathematical coupling. Am J Orthod Dentofacial Orthop 2006;130:62-8.

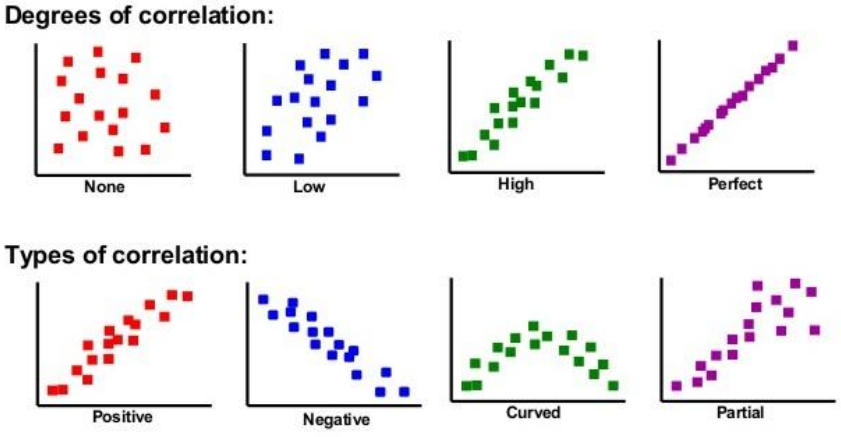

Figure I- The degree and types of correlation observed from scatter plot.

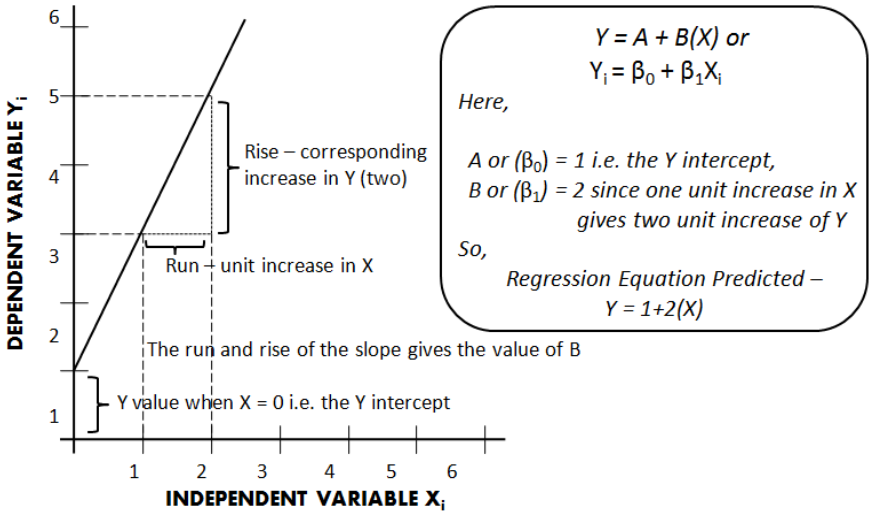

Figure II – Graph explaining the variables and the predicted regression equation.

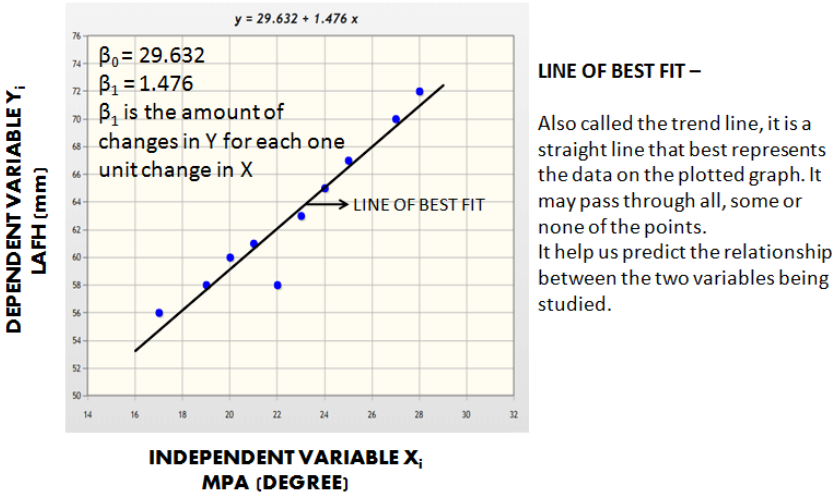

Figure III – Explaining the graph of the regression line.

S.No.	Mandibular Plane Angle (degree) X variable (Independent) (Xi)	Lower Anterior Facial Height (mm) Y variable (Dependent) (Yi)
1	17	56

2	21	61
3	24	65
4	20	60
5	27	70
6	22	58
7	23	63
8	19	58
9	25	67
10	28	72

Table I – The observed values of MPA and LAFH.

Regression Model	**Independent Variable**	**Dependent Variable**
1. Simple Linear Regression	Single	Single
	Continuous or Discrete	Only Continuous
	E.g. - Age	Blood Pressure
2. Multiple Linear Regression	Two or More	Single
	Continuous or Discrete	Only Continuous
	E.g. – Age, Sex and Weight	Blood Pressure

Table II - Explaining the simple and multiple linear regression models with examples.

Logistic Regression	**Independent Variable**	**Dependent Variable**
1. Simple	Single	Single
	Continuous or Discrete	Only Categorical
	E.g. – Lung cancer	Smokers and Non-smokers

2. Multiple	Two or More	Single
	Continuous or Discrete	Only Categorical
	E.g. – Lung cancer	Smokers and Non-smokers, Male and Females & Air pollution and Clean air

Table III - Explaining the simple and multiple logistic regressions with examples.

S. No.	Name of the Journal	Independent Variable (known)	Dependent Variable (calculated)	Regression Model used	Inference
1	Progress in Orthodontics 2014 [7]	Width of mandibular incisors	Width of unerupted canines and premolars	Simple Linear	By knowing the M-D width of the mandibular incisors the width of the unerupted canine and premolars can be calculated, thereby helping in space planning.
2	Eur J Orthod 1983 [8]	Form of roots	Likehood of root resorption	Simple Logitic	Curved, dilacerated, short and blunt roots are more susceptible to resorption
3	University of Iowa, 2015 [9]	Mandibular plane Angle	Facial height	Simple Linear	As the mandibular plane angle increases the facial height increases. The patient tends to grow vertically.
4	Progress in Orthodontics (2015) [10]	IGF-1 levels	Mandibular growth per year	Multiple Linear	A one unit increase in the IGF-1 change was associated with 0.00864 unit increase in mandibular length.
5	Am J Orthod Dentofacial Orthop 2013 [11]	Body Mass Index, age, sex, race	Dental age	Multiple Linear	An expected change in dental age for a unit change in BMI percentile is 0.005 year after adjusting for sex and age
6	American Journal of Orthodontics. 1974 [12]	Age of the patient, personality characterstics, attitude	Level of cooperation during treatment	Multiple Logistic	As the age of the patient increases the level of cooperation also increases owing to the maturation and development of communication skills.
7	Community Dent Oral Epidemiol 2016 [13]	Adverse birth outcomes	Childhood caries	Simple Linear	An inverse association between Low Birth Weight, Small for Gestational Age, preterm and childhood caries. with dental caries
8	Acta Dermatovenerol Croat. 2015 [14]	Stressful life events and Personality traits	Oral Lichen Planus	Simple Multiple	Anxiety, depression, familiar matters, war experiences are more commonly associated with OLP.

9	J Dent Res. 2016 [15]	Molar -Incisor Hypomineralization	Bone mass	Multiple Logistic	A low Bone Mineral Content (BMC) is associated with hypomineralized primary molars.
10	BMC Oral Health. 2015 [16]	Systolic BP, Diastolic BP and Hypertension	Sublingual varices	Multiple Logistic	An association was found between sublingual varices and hypertension.

Table IV - Examples showing situations where regression analysis has be done to explain the effects.