Linear regression analysis study

Article *in* Journal of the Practice of Cardiovascular Sciences · January 2018

DOI: 10.4103/jpcs.jpcs_8_18

2 authors:

Khushbu Kumari
All India Institute of Medical Sciences
5 PUBLICATIONS **327** CITATIONS
SEE PROFILE

Suniti Yadav
Indian Council of Medical Research
47 PUBLICATIONS **599** CITATIONS
SEE PROFILE

Curriculum in Cardiology

Statistics

Linear Regression Analysis Study

Khushbu Kumari, Suniti Yadav

Department of Anthropology, University of Delhi, New Delhi, India

Abstract

Linear regression is a statistical procedure for calculating the value of a dependent variable from an independent variable. Linear regression measures the association between two variables. It is a modeling technique where a dependent variable is predicted based on one or more independent variables. Linear regression analysis is the most widely used of all statistical techniques. This article explains the basic concepts and explains how we can do linear regression calculations in SPSS and excel.

Keywords: Continuous variable test, excel and SPSS analysis, linear regression

Introduction

The concept of linear regression was first proposed by Sir Francis Galton in 1894. Linear regression is a statistical test applied to a data set to define and quantify the relation between the considered variables. Univariate statistical tests such as Chi-square, Fisher's exact test, *t*-test, and analysis of variance (ANOVA) do not allow taking into account the effect of other covariates/confounders during analyses (Chang 2004). However, partial correlation and regression are the tests that allow the researcher to control the effect of confounders in the understanding of the relation between two variables (Chang 2003).

In biomedical or clinical research, the researcher often tries to understand or relate two or more independent (predictor) variables to predict an outcome or dependent variable. This may be understood as how the risk factors or the predictor variables or independent variables account for the prediction of the chance of a disease occurrence, i.e., dependent variable. Risk factors (or dependent variables) associate with biological (such as age and gender), physical (such as body mass index and blood pressure [BP]), or lifestyle (such as smoking and alcohol consumption) variables with the disease. Both correlation and regression provide this opportunity to understand the "risk factors-disease" relationship (Gaddis and Gaddis 1990). While correlation provides a quantitative way of measuring the degree or strength of a relation between two variables, regression analysis mathematically describes this relationship. Regression analysis allows predicting the value of a dependent variable based on the value of at least one independent variable.

In correlation analysis, the correlation coefficient "*r*" is a dimensionless number whose value ranges from −1 to +1. A value toward −1 indicates inverse or negative relationship, whereas towards +1 indicate a positive relation. When there is a normal distribution, the Pearson's correlation is used, whereas, in nonnormally distributed data, Spearman's rank correlation is used.

The linear regression analysis uses the mathematical equation, i.e., y = mx + c, that describes the line of best fit for the relationship between *y* (dependent variable) and *x* (independent variable). The regression coefficient, i.e., r^2 implies the degree of variability of *y* due to *x*.[1-8]

Significance of Linear Regression

The use of linear regression model is important for the following reasons:

a. Descriptive – It helps in analyzing the strength of the association between the outcome (dependent variable) and predictor variables
b. Adjustment – It adjusts for the effect of covariates or the confounders

Address for correspondence: Khushbu Kumari, Department of Anthropology, University of Delhi, New Delhi, India.
E-mail: khushukumari38@gmail.com

Access this article online

Quick Response Code:

Website: www.j-pcs.org

DOI: 10.4103/jpcs.jpcs_8_18

How to cite this article: Kumari K, Yadav S. Linear regression analysis study. J Pract Cardiovasc Sci 2018;4:33-6.

Table 1: SPSS table

Variables entered/removed[b]

Model	Variables entered	Variables removed	Method
I	Age (years)[a]		Enter

[a]All requested variables entered. [b]Dependent variable: Systolic blood pressure (mmHg)

Table 2: SPSS output with R^2

Model summary

Model	R	R^2	Adjusted R^2	Std. error of the estimate
I	0.696[a]	0.485	0.475	9.10072

[a]Predictors: (constant), age (years)

Table 3: Analysis of variance with P

ANOVA[a]

Model	Sum of squares	df	Mean square	F	Sig.
I Regression	4128.118	1	4128.118	49.843	0.000[b]
Residual	4389.628	53	82.823		
Total	8517.745	54	-		

[a]Predictors: (Constant), Age (years). [b]Dependent variable: Systolic blood pressure (mmHg)

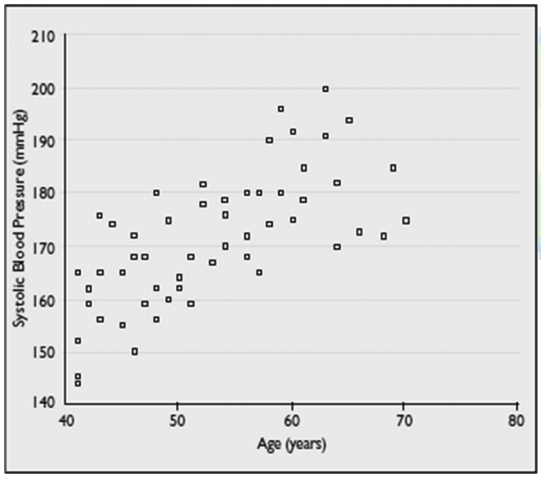

Figure 1: Scatter plot of systolic blood pressure versus age.

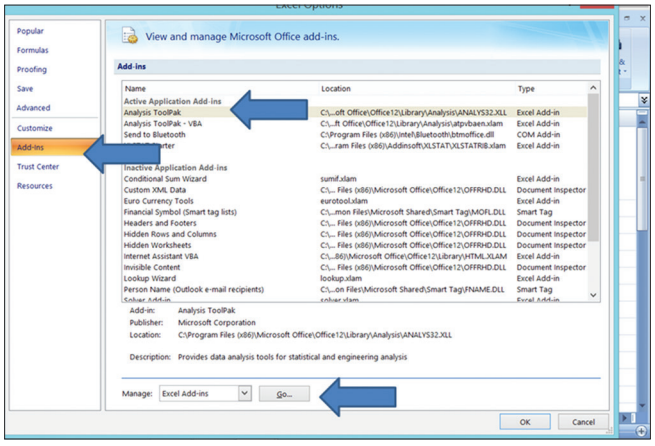

Figure 3: The Tool Pak. Choose Add Ins > Choose Analysis ToolPak and select Go.

c. Predictors – It helps in estimating the important risk factors that affect the dependent variable
d. Extent of prediction – It helps in analyzing the extent of change in the independent variable by one "unit" would affect the dependent variable
e. Prediction – It helps in quantifying the new cases.

Assumptions for Linear Regression

The underlying assumptions for linear regression are:

a. The values of independent variable "x" are set by the researcher
b. The independent variable "x" should be measured without any experimental error
c. For each value of "x," there is a subpopulation of "y" variables that are normally distributed up and down the Y-axis [Figure 1]
d. The variances of the subpopulations of "y" are homogeneous
e. The mean values of the subpopulations of "y" lie on a straight line, thus implying the assumption that there exists

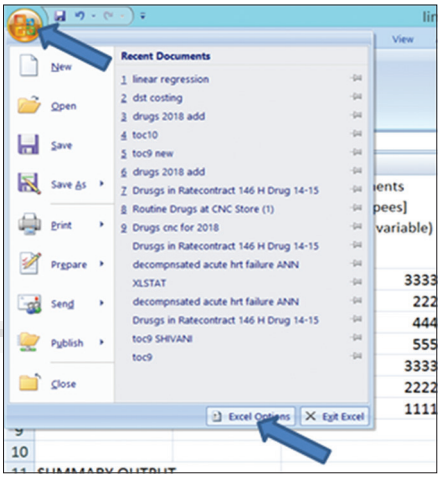

Figure 2: Starting Data Analysis ToolPak. Click the OFFICE button and choose Excel options.

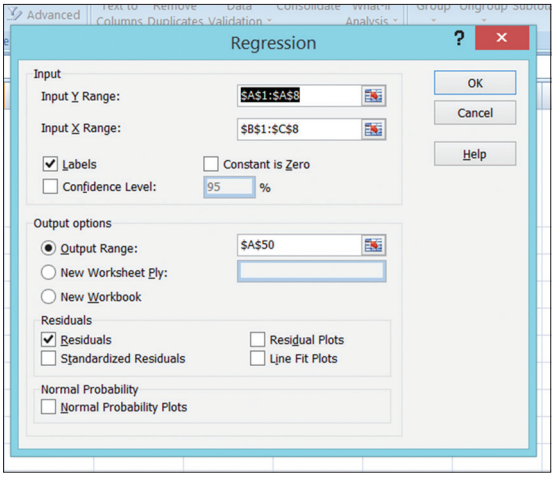

Figure 4: The regression screen. Choose Data > Data Analysis > Regression. Input y Range: A1:A8. Input X Range: B1:C8. Check Labels, Residuals, Output Range as A50.

Table 4: SPSS equation variables

Coefficients[a]							
Model	Unstandardised coefficients		Standardised coefficients	*t*	Sig.	95% Confidence interval for B	
	B	Std. error	Beta			Lower bound	Upper bound
I (Constant)	115.706	7.999	0.696	14.465	0.000	99.662	131.749
Age (years)	1.051	0.149		7.060	0.000	0.752	1.350

[a]Dependent variable: Systolic blood pressure (mmHg)

Table 5: Excel data set

A	B	C
Sale of medicine (number of units) (dependent variable)	Price (rupees per tablet) (independent variable)	TV advertisements budget (rupees) (independent variable)
8888	22	33,333
4444	55	2222
5555	33	4444
7777	22	5555
6666	55	33,333
7777	33	22,222
5555	44	11,111

Testing whether sale of medicine depends on price and advertisements

Table 6: Summary output

Regression statistics	Values	Explanation
Multiple *R*	0.96332715	Correlation coefficient: 1 means perfect correlation and 0 means none
R^2	0.927999198	Coefficient of determination: How many points fall on the regression line. Here, 92% points fall within the line
Adjusted R^2	0.891998797	Adjusted R^2: Adjusts for multiple variables, use with multiple variables
SE	516.3490153	
Observations	7	

a linear relation between the dependent and the independent variables

f. All the values of "y" are independent from each other, though dependent on "x."

Coefficient of Determination, R^2

The coefficient of determination is the portion of the total variation in the dependent variable that can be explained by variation in the independent variable(s). When R^2 is + 1, there exists a perfect linear relationship between x and y, i.e., 100% of the variation in y is explained by variation in x. When it is $0< R^2 <1$, there is a weaker linear relationship between x and y, i.e., some, but not all of the variation in y is explained by variation in x.

Linear Regression in Biological Data Analysis

In biological or medical data, linear regression is often used to describe relationships between two variables or among several variables through statistical estimation. For example, to know whether the likelihood of having high systolic BP (SBP) is influenced by factors such as age and weight, linear regression would be used. The variable to be explained, i.e., SBP is called the dependent variable, or alternatively, the response variables that explain it age, weight, and sex are called independent variables.

How to Calculate Linear Regression?

Linear regression can be tested through the SPSS statistical software (IBM Corp. Released 2011. IBM SPSS Statistics for Windows, Version 20.0. Armonk, NY: IBM Corp.) in five steps to analyze data using linear regression. Following is the procedure followed Tables 1-4:

Click Analyze > Regression > Linear > then select Dependent and Independent variable > OK (enter).

Example 1 – Data (n = 55) on the age and the SBP were collected and linear regression model would be tested to predict BP with age. After checking the normality assumptions for both variables, bivariate correlation is tested (Pearson's correlation = 0.696, $P < 0.001$) and a graphical scatter plot is helpful in that case [Figure 2].

Now to check the linear regression, put SBP as the dependent and age as the Independent variable.

This indicates the dependent and independent variables included in the test.

Pearson's correlation between SBP and age is given ($r = 0.696$). R^2 = 0.485 which implies that only 48.5% of the SBP is explained by the age of a person.

The ANOVA table shows the "usefulness" of the linear regression model with $P < 0.05$.

This provides the quantification of the relationship between age and SBP. With every increase of 1 year in age, the SBP (on the average) increases by 1.051 (95% confidence interval 0.752–1.350) units, $P < 0.001$. The constant here has no "practical" meaning as it gives the value of the SBP when age = 0.

Further, if more than one independent variable is added, the linear regression model would adjust for the effect of other dependent variables when testing the effect of one variable.

Example 2 – If we want to see the genetic effect of variables, i.e., the effect of increase in per allele dose of any genetic variant (mutation) on the disease or phenotype, linear

Table 7: Analysis of variance

	df	SS	MS	*F*	Significance *F*
Regression	2	13,745,386.778	6,872,693.389	25.777	0.005
Residual	4	1,066,465.222	266,616.306		
Total	6	14,811,852.000			
	Coefficients	**SE**	***t* statistic**		***P***
Intercept	8533.99	661.21	12.91		0.000
Price	−79.98	15.08	−5.31		0.006
TV adds	0.07	0.02	4.59		0.010
Equation=Sales: 8533.9 − 79.9 (price) + 0.07 (TV adds)					

If we now look at the last three columns, we can create an equation: Equation = Sales: 8533.9 − 79.9 (price) + 0.07 (TV ads). Therefore as the price goes up, the sales go down ($P = 0.006$), and the addition of TV ads is less (0.07 multiplied by advertisement money with a P value of 0.01). SE: Standard error, MS: Mean square, SS: Sum of square

regression is used in a similar way as described above. The three genotypes, i.e., normal homozygote AA, heterozygote AB and homozygote mutant BB may be coded as 1, 2, and 3, respectively. The test may be preceded, and in a similar way, the unstandardized coefficient (β) would explain the effect on the dependent variable with per allele dose increase.

Example 3 – Using Excel to see the relationship between sale of medicine with the price of the medicine and TV advertisements.

Table 5 contains data which can be entered into an Excel sheet. Follow instructions as shown in figures 2-4.

As shown in Table 6, Multiple R is the Correlation Coefficient, where 1 means a perfect correlation and zero means none. R Square is the coefficient of determination which here means that 92% of the variation can be explained by the variables. Adjusted R square adjusts for multiple variables and should be used here. here. Table 7 shows how to create a linear regression equation from the data.

Conclusion

The techniques for testing the relationship between two variables are correlation and linear regression. Correlation quantifies the strength of the linear relationship between a pair of variables, whereas regression expresses the relationship in the form of an equation. In this article, we have used simple examples and SPSS and excel to illustrate linear regression analysis and encourage the readers to analyze their data by these techniques.

Financial support and sponsorship

Nil.

Conflicts of interest

There are no conflicts of interest.

References

1. Schneider A, Hommel G, Blettner M. Linear regression analysis: Part 14 of a series on evaluation of scientific publications. Dtsch Arztebl Int 2010;107:776-82.
2. Freedman DA. Statistical Models: Theory and Practice. Cambridge, USA: Cambridge University Press; 2009.
3. Chan YH. Biostatistics 201: Linear regression analysis. Age (years). Singapore Med J 2004;45:55-61.
4. Chan YH. Biostatistics 103: Qualitative data – Tests of independence. Singapore Med J 2003;44:498-503.
5. Gaddis ML, Gaddis GM. Introduction to biostatistics: Part 6, correlation and regression. Ann Emerg Med 1990;19:1462-8.
6. Mendenhall W, Sincich T. Statistics for Engineering and the Sciences. 3rd ed. New York: Dellen Publishing Co.; 1992.
7. Panchenko D. 18.443 Statistics for Applications, Section 14, Simple Linear Regression. Massachusetts Institute of Technology: MIT OpenCourseWare; 2006.
8. Elazar JP. Multiple Regression in Behavioral Research: Explanation and Prediction. 2nd ed. New York: Holt, Rinehart and Winston; 1982.