J Bras Pneumol. 2020;46(1):e20200030
https://dx.doi.org/10.1590/1806-3713/e20200030

Correlation vs. regression in association studies

Suzana Erico Tanni[1,2], Cecilia Maria Patino[1,3], Juliana Carvalho Ferreira[1,4]

When the goal of a researcher is to evaluate the relationship between variables, both correlation and regression analyses are commonly used in medical science. Although related, correlation and regression are not synonyms, and each statistical approach is used for a specific purpose and is based on a set of specific assumptions.

When testing the correlation between two variables, we use the correlation coefficient (r) to quantify both the strength and the direction of the relationship between two numeric variables, the results ranging from −1 to 1. When r = 0, this indicates that there is no linear relationship between the two variables; when r = 1, this indicates a perfect positive relationship between the two variables and implies that as the value of one variable increases, the value of the other one also increases (Figure 1). When r = −1, this indicates a perfect negative relationship and implies that as the value of one variable increases, the value of the other one decreases. In most cases, the strength of the relationship between the variables is not perfect; therefore, r is not exactly 1 or −1. The strength of a correlation is commonly interpreted as weak (r < ±0.4), moderate (r ranging from ±0.4 to ±0.7), and strong (r > ±0.7).[1] Lastly, we highlight that when correlation is used as a statistical approach, the data should be derived from a random sample; the variables should be continuous; the data should not include outliers; each pair of variables need to be independent[1]; and the correlation does not necessarily imply a cause-and-effect relationship.

Regression is indicated when one of the variables is an outcome and the other one is a potential predictor of that outcome, in a cause-and-effect relationship. If the outcome is a continuous variable, a linear regression model is indicated, and, if it is binary, a logistic regression is used. Regression also quantifies the direction and strength of the relationship between two numeric variables, X (the predictor) and Y (the outcome); however, in contrast with correlation, these two variables are not interchangeable, and correctly identifying the outcome and the predictor is key. Regression models additionally permit the evaluation of more than one predictor variable, another important difference from correlation analysis.[2]

Regression is a linear mathematical model represented by the equation $Y = \beta_0 + \beta_1 X$ (Figure 1). When the value of X (the predictor) is zero, the value of Y is β_0 (the line intercept), and β_1 is the slope, which gives us information of the magnitude and direction of the association between X and Y, similarly to the correlation coefficient. When $\beta_1 = 0$, there is no association between X and Y. When $\beta_1 > 0$ or $\beta_1 < 0$, the association between X and Y is positive or negative, respectively. Important assumptions of linear regression are normality and linearity of the outcome variable, independence between the two variables, and equal variance of the outcome variable across the regression line.[2]

In conclusion, when evaluating the relationship between two variables, we need to understand the differences between correlation and regression and choose which statistical test is better to answer the research question.

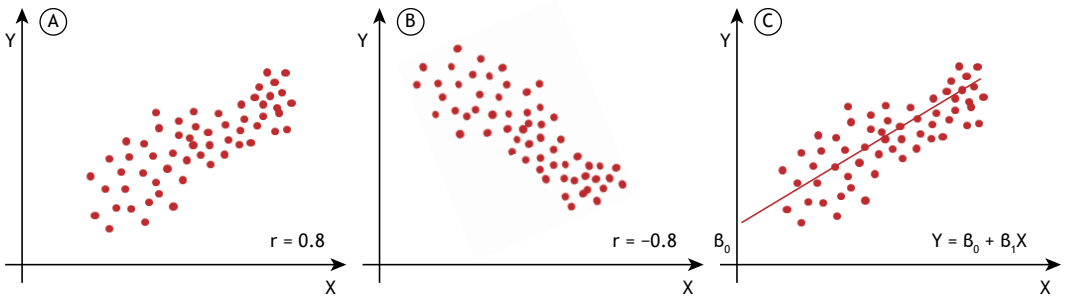

Figure 1. Scatter plots with simulated values of two variables, X and Y. In A, the circles represent pairs of simulated variables X and Y, showing that increases in X are associated with increases in Y: correlation coefficient (r) = 0.8. In B, the circles represent pairs of simulated variables X and Y, showing that increases in X are associated with decreases in Y: r = −0.8. In C, the circles represent the same pairs of simulated values of variables X and Y shown in A, fitted with a linear regression model, in which β_0 is the intercept and β_1 is the slope of the curve.

REFERENCES

1. Schober P, Boer C, Schwarte LA. Correlation Coefficients: Appropriate Use and Interpretation. Anesth Analg. 2018;126(5):1763–1768. http://doi:10.1213/ANE.0000000000002864
2. Kutner MH, Nachtsheim CJ, Neter J, Li W. Simple Linear Regression. In: Kutner MH, Nachtsheim CJ, Neter J, Li W. Applied linear statistical models. 5th ed. New York: McGraw-Hill; 2005. p. 1-87.

1. Methods in Epidemiologic, Clinical, and Operations Research–MECOR–program, American Thoracic Society/*Asociación Latinoamericana del Tórax*, Montevideo, Uruguay.
2. Departamento de Medicina Interna, Área de Pneumologia, Faculdade de Medicina de Botucatu, Universidade Estadual Paulista – UNESP – Botucatu (SP) Brasil.
3. Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, CA, USA.
4. Divisão de Pneumologia, Instituto do Coração, Hospital das Clínicas, Faculdade de Medicina, Universidade de São Paulo, São Paulo (SP) Brasil.

BY-NC © 2020 Sociedade Brasileira de Pneumologia e Tisiologia