Common pitfalls in statistical analysis: Linear regression analysis

Rakesh Aggarwal, Priya Ranganathan[1]

Department of Gastroenterology, Sanjay Gandhi Postgraduate Institute of Medical Sciences, Lucknow, Uttar Pradesh, [1]Department of Anaesthesiology, Tata Memorial Centre, Mumbai, Maharashtra, India

Abstract

In a previous article in this series, we explained correlation analysis which describes the strength of relationship between two continuous variables. In this article, we deal with linear regression analysis which predicts the value of one continuous variable from another. We also discuss the assumptions and pitfalls associated with this analysis.

Keywords: Biostatistics, linear model, regression analysis

Address for correspondence:
Dr. Priya Ranganathan, Department of Anaesthesiology, Tata Memorial Centre, Ernest Borges Road, Parel, Mumbai - 400 012, Maharashtra, India.
E-mail: drpriyaranganathan@gmail.com

We often have information on two numeric characteristics for each member of a group and believe that these are related to each other – i.e. values of one characteristic vary depending on the values of the other. For instance, in a recent study, researchers had data on body mass index (BMI) and mid-upper arm circumference (MUAC) on 1373 hospitalized patients, and they decided to determine whether there was a relationship between BMI and MUAC.[1] In such a situation, as we discussed in a recent piece on "Correlation" in this series,[2] the researchers would plot the data on a scatter diagram. If the dots fall roughly along a straight line, sloping either upwards or downwards, they would conclude that a relationship exists. As a next step, they may be tempted to ask whether, knowing the value of one variable (MUAC), it is possible to predict the value of the other variable (BMI) in the study group. This can be done using "simple linear regression" analysis, also sometimes referred to as "linear regression." The variable whose value is known (MUAC here) is referred to as the independent (or predictor or explanatory) variable, and the variable whose value is being predicted (BMI here) is referred to as the dependent (or outcome or response) variable. The independent and dependent variables are, by convention, referred to as "x" and "y" and are plotted on horizontal and vertical axes, respectively.

At times, one is interested in predicting the value of a numerical response variable based on the values of more than one numeric predictors. For instance, one study found that whole-body fat content in men could be predicted using information on thigh circumference, triceps and thigh skinfold thickness, biceps muscle thickness, weight, and height.[3] This is done using "multiple linear regression." We will not discuss this more complex form of regression.

Although the concepts of "correlation" and "linear regression" are somewhat related and share some assumptions, these also have some important differences, as we discuss later in this piece.

THE REGRESSION LINE

Linear regression analysis of observations on two variables (x and y) in a sample can be looked upon as plotting the data and drawing a best fit line through these. This "best fit" line is so chosen that the sum of squares of all the residuals (the vertical distance of each point from the line) is a minimum – the so-called "least squares line" [Figure 1].

Access this article online	
Quick Response Code:	Website: www.picronline.org
	DOI: 10.4103/2229-3485.203040

How to cite this article: Aggarwal R, Ranganathan P. Common pitfalls in statistical analysis: Linear regression analysis. Perspect Clin Res 2017;8:100-2.

This line can be mathematically defined by an equation of the form:

$$Y = a + bx$$

Where "x" is the known value of independent (or predictor or explanatory) variable, "Y" is the predicted (or fitted) value of "y" (dependent, outcome, or response variable) for the given value of "x", "a" is called as the "intercept" of the estimated line and represents the value of Y when x = 0, and "b" is called as the "slope" of the estimated line and represents the amount by which Y changes on average as "x" increases by one unit. It is also referred to as "coefficient," "regression coefficient," or "gradient." Note that lowercase letters (x and y) are used to denote the actual values and capital letters (Y) for predicted values.

The value of "b" is positive when the value of Y increases with each unit increase in x and is negative if the value of Y decreases with each unit increase in x [Figure 2]. If the value of Y does not change with x, the value of "b" would be expected to be 0. Furthermore, the higher the magnitude of "b," the steeper is the change in Y with change in x.

In the example of BMI and MUAC,[1] the linear correlation equation was: BMI = –0.042 + 0.972 × MUAC (in cm). Here, +0.972 is the slope or coefficient and indicates that, on average, BMI is expected to be higher by 0.972 units for each unit (cm) increase in MUAC. The first term in the equation (i.e., –0.042) represents the intercept and would be the expected BMI if a person had MUAC of 0 (a zero or negative value of BMI may appear unusual but more on this later).

ASSUMPTIONS

Regression analysis makes several assumptions, which are quite akin to those for correlation analysis, as we discussed in a recent issue of the journal.[1] To recapitulate, first, the relationship between x and y should be linear. Second, all the observations in a sample must be independent of each other; thus, this method should not be used if the data include more than one observation on any individual. Furthermore, the data must not include one or a few extreme values since these may create a false sense of relationship in the data even when none exists. If these assumptions are not met, the results of linear regression analysis may be misleading.

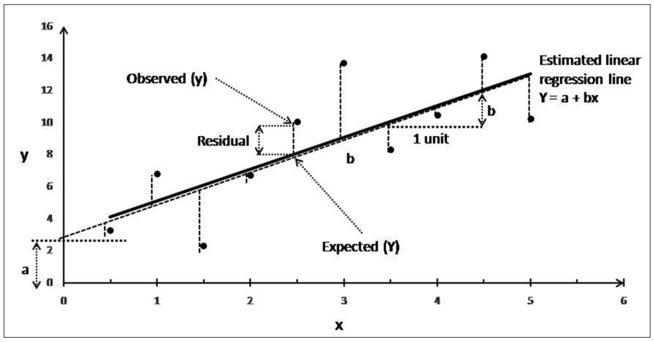

Figure 1: Data from a sample and estimated linear regression line for these data. Each dot corresponds to a data point, i.e., an individual pair of values for x and y, and the vertical dashed lines from each dot represent residuals. The capital letters (Y) are used to indicate predicted values and lowercase letters (x and y) for known values. Intercept is shown as "a" and slope or regression coefficient as "b"

CORRELATION VERSUS REGRESSION

Correlation and regression analyses are similar in that these assess the linear relationship between two quantitative variables. However, these look at different aspects of this relationship. Simple linear regression (i.e., its coefficient or "b") predicts the nature of the association – it provides a means of predicting the value of dependent variable using the value of predictor variable. It indicates how much and in which direction the dependent variable changes on average for a unit increase in the latter. By contrast, correlation (i.e., correlation coefficient or "r") provides a measure of the strength of linear association – a measure of how closely the individual data points lie on the regression line. The values of "b" and "r" always carry the same sign – either both are positive or both are negative. However, their magnitudes can vary widely. For the same value of "b," the magnitude of "r" can vary from 1.0 to close to 0.

ADDITIONAL CONSIDERATIONS

Some points must be kept in mind when interpreting the results of regression analysis. The absolute value of regression

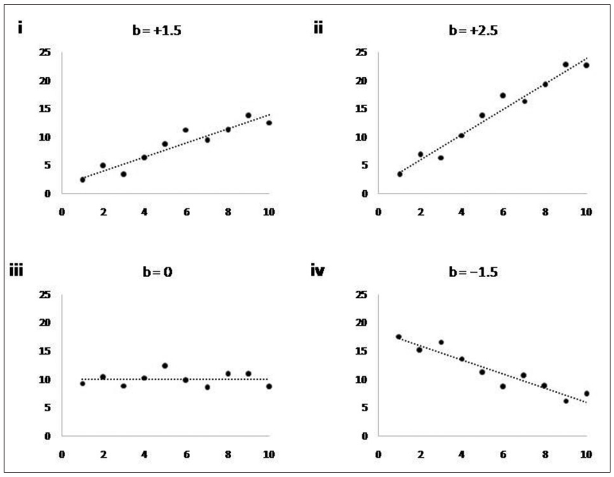

Figure 2: Relationships between two quantitative variables and their regression coefficients ("b"). "b" represents predicted change in the value of dependent variable (on Y axis) for each one unit increase in the value of independent variable (on X axis). "b" is positive, zero, or negative, depending on whether, as the independent variable increases, the value of dependent variable is predicted to increase (panels i and ii), remain unchanged (iii), or decrease (iv). A higher absolute value of "b" indicates that the independent variable changes more for each unit increase in the predictor (ii vs i)

coefficient ("b") depends on the units used to measure the two variables. For instance, in a linear regression equation of BMI (independent) versus MUAC (dependent), the value of "b" will be 2.54-fold higher if the MUAC is expressed in inches instead of in centimeters (1 inch = 2.54 cm); alternatively, if the MUAC is expressed in millimeters, the regression coefficient will become one-tenth of the original value (1 mm = 1/10 cm). A change in the unit of "y" will also lead to a change in the value of the regression coefficient. This must be kept in mind when interpreting the absolute value of a regression coefficient.

Similarly, the value of "intercept" also depends on the unit used to measure the dependent variable. Another important point to remember about the "intercept" is that its value may not be biologically or clinically interpretable. For instance, in the MUAC-BMI example above, the intercept was −0.042, a negative value for BMI which is clearly implausible. This happens when, in real-life, the value of independent variable cannot be 0 as was the case for the MUAC-BMI example above (think of MUAC = 0; it simply cannot occur in real-life). Furthermore, a regression equation should be used for prediction only for those values of the independent variable that lie within in the range of the latter's values in the data originally used to develop the regression equation.

Financial support and sponsorship

Nil.

Conflicts of interest

There are no conflicts of interest.

REFERENCES

1. Benítez Brito N, Suárez Llanos JP, Fuentes Ferrer M, Oliva García JG, Delgado Brito I, Pereyra-García Castro F, *et al.* Relationship between mid-upper arm circumference and body mass index in inpatients. PLoS One 2016;11:e0160480.
2. Aggarwal R, Ranganathan P. Common pitfalls in statistical analysis: The use of correlation techniques. Perspect Clin Res 2016;7:187-90.
3. Bielemann RM, Gonzalez MC, Barbosa-Silva TG, Orlandi SP, Xavier MO, Bergmann RB, *et al.* Estimation of body fat in adults using a portable A-mode ultrasound. Nutrition 2016;32:441-6.