ENDGAMES

STATISTICAL QUESTION

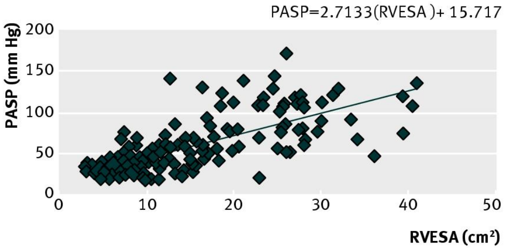
Simple linear regression

Philip Sedgwick *reader in medical statistics and medical education*

Centre for Medical and Healthcare Education, St George’s, University of London, Tooting, London, UK

Researchers investigated the association of right ventricular size and function with varying degrees of pulmonary hypertension. A cross sectional study design was used. Participants were 190 patients referred to a pulmonary hypertension clinic.[1]

Measurements of right ventricular size included right ventricular end systolic area (RVESA) recorded echocardiographically. The extent of pulmonary hypertension was indicated by pulmonary artery systolic pressure (PASP). A scatter plot of pulmonary artery systolic pressure against right ventricular end systolic area was presented (figure). Linear regression analysis was used to examine the association between right ventricular size and degree of pulmonary hypertension. The resulting fitted linear regression line was given by PASP=2.7133RVESA+15.717.



Scatter plot of pulmonary artery systolic pressure (PASP) against right ventricular end systolic area (RVESA)

Which of the following statements, if any, are true for the linear regression line?

a) Pulmonary artery systolic pressure can be predicted given a value of right ventricular end systolic area

b) Right ventricular end systolic area can be predicted given a value of pulmonary artery systolic pressure

c) It was assumed that the variation in pulmonary artery systolic pressure was equal for all values of right ventricular end systolic area

d) The line can be extrapolated outside the observed range of values for right ventricular end systolic area

Answers

Statements *a* and *c* are true, whereas *b* and *d* are false.

The researchers investigated the association between right ventricular end systolic area and severity of pulmonary hypertension. The scatter plot suggested that a linear association existed—as the right ventricular end systolic area increased so did the pulmonary artery systolic pressure. This association was investigated using simple linear regression, often referred to as linear regression, which quantifies the nature of any linear association between two variables. The linear association is described by a mathematical equation.

Pulmonary artery systolic pressure was thought to be dependent on, or at least associated with, right ventricular end systolic area. A change in right ventricular size was associated with a change in pulmonary artery systolic pressure, rather than the other way round. The straight line that best describes the association between the two variables was calculated as PASP=2.7133(RVESA)+15.717; it describes how much pulmonary artery systolic pressure (PASP) changes on average as right ventricular end systolic area (RVESA) changes. The regression line is referred to as the regression of pulmonary artery systolic pressure on right ventricular end systolic area. PASP is termed the dependent variable and RVESA the independent, predictor, or explanatory variable. The intercept of the regression line—the value of PASP when the line crosses the Y axis—is calculated by setting RVESA to zero and is given by 15.717. The slope or gradient of the regression line is 2.7133—the coefficient of RVESA in the equation; this represents the amount by which PASP changes on average for each unit (1 cm^2) increase in RVESA. Therefore, for each value of RVESA the predicted PASP can be calculated.

The linear regression line is calculated using the so called method of ordinary least squares, often called least squares. This method involves consideration of all possible straight lines through the points on the scatter plot. For each possible straight line, the residuals—the vertical differences between each point on the scatter plot and the straight line—are derived. A residual is therefore the difference between the observed value of pulmonary artery systolic pressure for a patient and the predicted

p.sedgwick@sgul.ac.uk

value given the patient's measurement of right ventricular end systolic area. The residuals are measured on the same scale as pulmonary artery systolic pressure. The resulting fitted regression line was such that it was the one of best fit—that is, the sum of the squared residuals was minimised across all possible lines.

In the example above, the linear regression line was the regression of pulmonary artery systolic pressure on right ventricular end systolic area. It can be used to predict pulmonary artery systolic pressure for a given value of right ventricular end systolic area (*a* is true). It is not possible, however, to predict right ventricular end systolic area for a given artery systolic pressure (*b* is false). This is because the least squares regression line was based on the residuals for pulmonary artery systolic pressure. To predict right ventricular end systolic area from a given pulmonary artery systolic pressure, right ventricular end systolic area would need to be regressed on pulmonary artery systolic pressure, which would involve the residuals for right ventricular end systolic area.

The application of linear regression analysis made a series of assumptions. These included, perhaps obviously, that there was a linear relation between pulmonary artery systolic pressure and right ventricular end systolic area. Inspection of the scatter plot (figure) suggested that such an association existed. The second assumption was that the observations on the scatter plot are independent of each other—that is, each patient had only one observation of pulmonary artery systolic pressure and right ventricular end systolic area. Thirdly, it was assumed that the residuals were normally distributed; this could be verified by inspection of a histogram. It was also assumed that the variation in the distribution of pulmonary artery systolic pressure was the same for all values of right ventricular end systolic area (*c* is true). Inspection of the scatter plot (figure) suggests that this assumption is not true because as right ventricular end systolic area increases, the observed pulmonary artery systolic pressure measurements increase in variation around the fitted linear regression line. A transformation of the dependent variable—pulmonary artery systolic pressure—might therefore be considered. A logarithmic transformation, described in a previous question,[2] might achieve constant variation in the distribution of pulmonary artery systolic pressure for all values of right ventricular end systolic area.

The linear regression line can be used to predict pulmonary artery systolic pressure only for the observed range of values for right ventricular end systolic area in the sample (*d* is false). For example, it is not possible to predict pulmonary artery systolic pressure when right ventricular end systolic area equals 50 cm^2; this value was outside the range of observed values for the independent variable. It is not possible to predict the association between pulmonary artery systolic pressure and right ventricular end systolic area for values of right ventricular end systolic area that were not observed.

The purposes of linear regression and correlation are often confused. Correlation has been described in previous questions,[3 4] Linear regression and correlation will be compared and contrasted in a future statistical question.

In the example above, the application of simple linear regression predicted pulmonary artery systolic pressure from only one explanatory variable—right ventricular end systolic area. Multiple linear regression analysis is a natural extension of simple linear regression with the inclusion of more than one explanatory variable. Multiple linear regression will be discussed in a future statistical question.

Competing interests: None declared.

[1] López-Candales A, Dohi K, Rajagopalan N, Edelman K, Gulyasy B, Bazaz R. Defining normal variables of right ventricular size and function in pulmonary hypertension: an echocardiographic study. *Postgrad Med J* 2008;84:40-5.

[2] Sedgwick P. Log transformation of data. *BMJ* 2012;345:e6727.

[3] Sedgwick P. Pearson's correlation coefficient. *BMJ* 2012;345:e4483.

[4] Sedgwick P. Correlation. *BMJ* 2012;345:e5407.

Cite this as: *BMJ* 2013;346:f2340