Advanced Statistics for Exotic Animal Practitioners

John Hodsoll, PhD*, Jennifer M. Hellier, BSc (Hons), Elizabeth G. Ryan, PhD

KEYWORDS

• Statistics • Exotic animals • Correlation • Regression

KEY POINTS

- Correlation estimates the strength and direction of an association between 2 random variables.
- Linear regression aims to quantify and predict the values of a continuous random variable based on the values of an explanatory variable or variables.
- Logistic regression extends linear models to binary outcomes.
- Statistical models rely on assumptions which should be critically evaluated.

VISUALISING THE RELATIONSHIP BETWEEN 2 VARIABLES

Scatter plots are an important graphical method to understand the nature of the association between 2 random variables. In a scatter plot, x and y coordinate points on a graph are determined by the numerical value of the variables. If there is no correlation between the variables, the points of the variables are scattered randomly across the plane. In contrast, if there is a linear correlation, the distribution of points will follow a straight line and be scattered randomly above and below the line, with most data close to the line. The strength of the association depends on the spread of the data around the line. An apparently strong linear association is shown in **Fig. 1**, which shows a scatter plot for the heart weight of 143 cats weighing more than 2 kg against their body weight. The line here represents the best-fit line through the data, which approximates the overall trend of the data most completely. There are various distance measures used to calculate this line. Most typically the least squares approach is used, as in linear regression (discussed later).

Disclosure: Dr J. Hodsoll receives salary support from the National Institute for Health Research (NIHR) Mental Health Biomedical Research Center at South London and Maudsley NHS Foundation Trust and King's College London. This article is independent research part funded by the National Institute for Health Research (NIHR) Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King's College London. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health.
Department of Biostatistics, Institute of Psychiatry, Psychology & Neuroscience, King's College London, 16 De Crespeigny Park, London SE5 8AF, UK
* Corresponding author.
E-mail address: john.hodsoll@kcl.ac.uk

Vet Clin Exot Anim 20 (2017) 961–972
http://dx.doi.org/10.1016/j.cvex.2017.04.015 **vetexotic.theclinics.com**

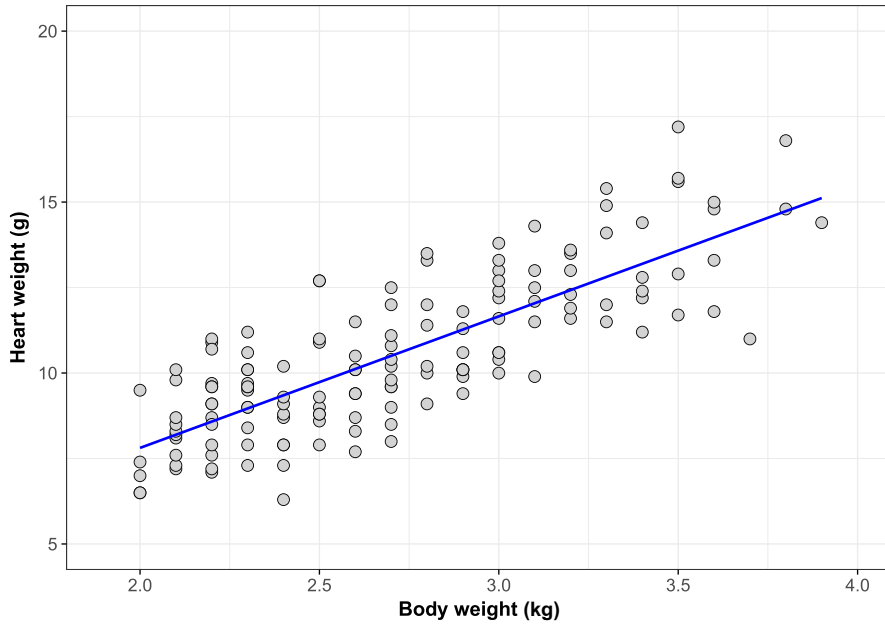

Fig. 1. Linear association between body and heart weight in a sample of adult cats weighing more than 2 kg.

The value of graphical analysis can be shown in **Fig. 2**, which shows 4 data sets created by the statistician Anscombe, each with 2 random variables X and Y.[1] Each of the variables in the 4 data sets have very similar descriptive statistics; the mean and standard deviation are the same, as is (of particular relevance) the linear line of best fit. However, the relationship between the 2 variables in the 4 data sets is very different and this becomes clear when the associations are inspected visually. Only example 2A is a good candidate for linear correlation or regression. For 2B, the relationship between X and Y is curvilinear or quadratic, set 2C has a large outlier, and for 2D there is no variability in X apart from 1 large observation.

CORRELATION

Correlation assesses the strength of an association between 2 random variables. For linear associations between 2 continuous variables, the appropriate statistic is the Pearson product moment correlation coefficient (commonly referred to as *r*). The value for *r* estimates how 2 random normally distributed variables X and Y covary with one another or how variations in one variable are related to variations in the other. For example, for a particular animal, whether a value exceeding the mean for one variable suggests a value higher than the mean for another variable. Alternatively, it may imply a value less than the mean or no systematic relationship. The correlation coefficient is calculated as the covariance divided by the standard deviations for the two variables, and ranges from −1 to 1. Variables are positively correlated if, when x increases, y also increases, or if, when x decreases, y also decreases. A negative correlation between 2 variables is identified when, as X increases, Y decreases, or as X decreases, Y increases. As such, calculation of *r* allows measurement of the extent to which knowing the value of X helps to predict the value of Y. It is scaled according to standard deviation, and the value is independent of the units of measurement for X and Y.