BJP British Journal of Pharmacology

DOI:10.1111/j.1476-5381.2012.01960.x
www.brjpharmacol.org

EDITORIAL

Categorized or continuous? Strength of an association – and linear regression

Gordon B Drummond[1] and Sarah L Vowler[2]

[1]*Department of Anaesthesia and Pain Medicine, University of Edinburgh, Royal Infirmary, Edinburgh, Edinburgh, UK, and* [2]*Cancer Research UK, Cambridge Research Institute, Li Ka Shing Centre, Cambridge, UK*

Correspondence
Gordon B. Drummond, Department of Anaesthesia and Pain Medicine, University of Edinburgh, Royal Infirmary, 51 Little France Crescent, Edinburgh, EH16 4HA, UK. E-mail: g.b.drummond@ed.ac.uk

This article is being published in *The Journal of Physiology*, *Experimental Physiology*, *the British Journal of Pharmacology*, *Advances in Physiology Education*, *Microcirculation*, and *Clinical and Experimental Pharmacology and Physiology*.

Gordon Drummond is Senior Statistics Editor for *The Journal of Physiology*.

Sarah Vowler is Senior Statistician in the Bioinformatics Core at Cancer Research UK's Cambridge Research Institute.

This article is the 8th in a series of articles on best practice in statistical reporting. All the articles can be found at http://onlinelibrary.wiley.com/journal/10.1111/(ISSN)1476-5381/homepage/statistical_reporting.htm.

Key points

- Correlation and regression are used with continuous variables.
- Plot the variables in correlation and regression relationships to aid interpretation.
- An association between two discrete measurements is assessed by correlation.
- Regression *describes* and *quantifies* a relationship between an independent factor and a dependent variable; *prediction* is also possible.
- Few biological relationships are truly linear.
- Regression can be distorted by outlying values.
- Absence of a linear regression does not mean a relationship is not present.
- Regression is very frequently misused and mis-applied.

Our previous article in this series considered ANOVA (Drummond and Vowler, 2012), which is applied to measurements made of samples from different groups. That approach considers the variation of the data, and specifically the variation *between* samples, which is related to the presence of different levels of the factor related to a group. The variation between samples is contrasted with the residual variation, found *within* the samples. We consider the possibility that the groups could be different, because of the different conditions of a factor. This is as far as the analysis can extend: the consideration is restricted to groups characterized by the different category of the factor being considered. For example, in Figure 1A, the factor we had considered in our samples of jumping frogs is the State from which they were sampled. The data were *categorized* according to the geographic origin of the frogs. However, in many biological experiments, the factor considered may not be just a simple category, but it may be expressed in terms of order, or even as a continuous variable. If this can be done, then other helpful and powerful methods of statistical analysis can be used.

Ranked tests

If the categories can be ranked, then tests can exploit this ranking. These include the Mann–Whitney (also known as the Wilcoxon rank sum) test, and Kendall's and Spearman's rank correlation tests, and simple examples are clearly described by Moses and colleagues (Moses *et al.*, 1984). By arranging the levels in a way that allows a factor to be logically graded (such as either not present, mild, moderate or severe), ordered levels are related to the measurements. For example, in Figure 1B, we have ranked the geographic origin from east to west, and there appears to be a possible association.

Quantitative variables

In many other experiments, we use quantitative variables rather than ranked categories, and relate the values observed

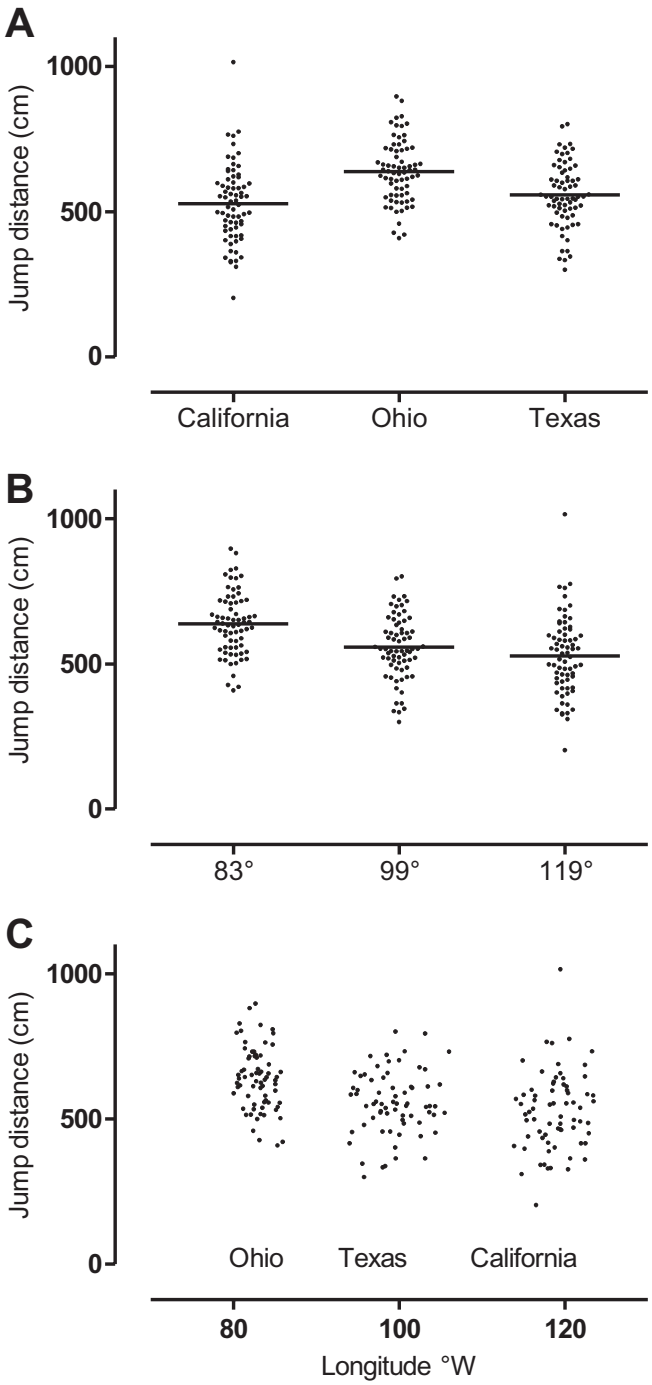

Figure 1

(A) The original samples, categorized by US State. (B) The samples from each State, ordered by longitude. (C) The individual frog jump lengths, related to the longitude of where the frog was found.

in one quantitative variable to another associated quantitative variable. The linkage, or *association*, between them can be mathematically determined by their *correlation*. When correlation is calculated, the strength of the association indicates how much of the variation of the two features occurs in the 'same direction'. A simple example might be body weight and height: if the correlation between these measures in a sample of adult females were perfect, the strength of the association, indicated by the correlation coefficient, would be 1, indicating that these two factors are exactly linked. The correlation coefficient can range from –1 (a perfect inverse association) to 1 (a perfect positive association). The mathematics of correlation is explained by Curran-Everett (Curran-Everett, 2010).

When using correlation procedures, a simple plot should be made and inspected first. In biology, truly linear associations are probably rare. Many interactions between factors may tend to a maximum, some are reciprocal, others could be logarithmic; forcing a straight line through such relationships is possible but illogical. The common correlation method used, Pearson's method, is only valid if at least one and preferably both of the measures are normally distributed. A dot plot of the association may show the distribution and prevent inappropriate use of the method. A dot plot can also prevent illogical conclusions. Figure 2 shows some of the patterns that may be found.

Correlation is generally only applicable if the variation in the measurements is uniform. Example B in Figure 2 shows a pattern that is not uncommon in biology, where the variability of the measures gets greater as the values increase. This feature (heteroscedascity) makes simple analysis unwise, although there are means to correct for it. In some instances (E and F), a straight line is not the most appropriate way to describe the link; in some, the relationship may be more complex (two different groups in panel G); and in others, the correlation may be the result of outlying points (panel H) that deserve careful consideration before conclusions are drawn.

Correlation is very frequently abused. If many variables are measured, and correlations sought, then some of them – often unrelated – will correlate, by chance. In particular, possibly false associations may be drawn from a time series, such as changes in the prevalence of obesity and the manufacture of fashion clothing over a number of years. A correlation may result from data that are biased: if for example we only measured the frogs we could catch easily, we might not appreciate that Ohio frogs were even better jumpers. We have already considered how the effects of another factor such as sex (a covariate) could skew the measurements: finding young female frogs and older males could cause a false association. Correlation is not necessarily the best or only means to assess agreement between two methods of measurement. Mathematical linkage between two measurements causes spurious correlation (Archie, 1981). Obvious links could be a change in weight in relation to starting weight, or relating a part to the whole. Others are more subtle: for example, when oxygen delivery and oxygen consumption are both calculated using the same measure of cardiac output and of arterial oxygen content, this generates a false association (Walsh and Lee, 1998).

'Equal' association in two measures, assessed by correlation, is not often sought in experimental situations. More often in the laboratory, measurements are considered to be affected by a factor that is not only quantifiable, but one that can be adjusted or predetermined, rather than randomly occurring. For this type of analysis, linear regression is often used. This should be approached as a separate statistical method from correlation, although it is often considered in the same chapter of the statistical textbooks.

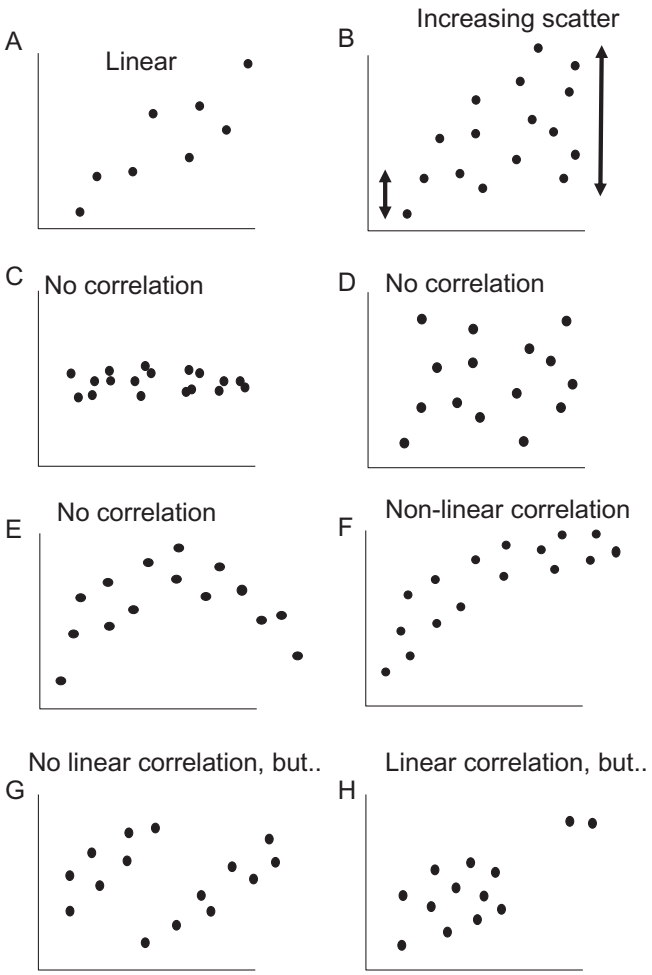

Figure 2

The variety of possible associations between two measures. (A) Positive correlation. (B) Correlation where the variation is not uniform: a common pattern in biological variables. (C) A plot where there is little variation in one measure: no correlation. (D) Scattered values: no correlation. (E) If an association is not linear, it may have no linear correlation. (F) Some associations may have a linear correlation, but are better described in other ways. (G) A dot plot may reveal unexpected complex associations. (H) A correlation can be caused by a few outlying values that have a strong effect.

Linear regression

What is linear regression? When we considered ANOVA, we attributed some of the variation in a measurement to factors that were classified as categories: the State where the frog was found, for example. In the theory of regression analysis, variation is attributed not to a specific category, but to an input factor that varies continuously. It aims to *describe* mathematically the association between the measured value and this input factor. Variation in the 'dependent' measure is *explained* in part by the magnitude of this input value, which is termed the 'independent' variable. In a simple regression analysis, the '% explained' is given by the square of the regression coefficient, R^2, expressed as a percentage. Thus if R^2 were 0.7, in a simple regression analysis we can say that 70%

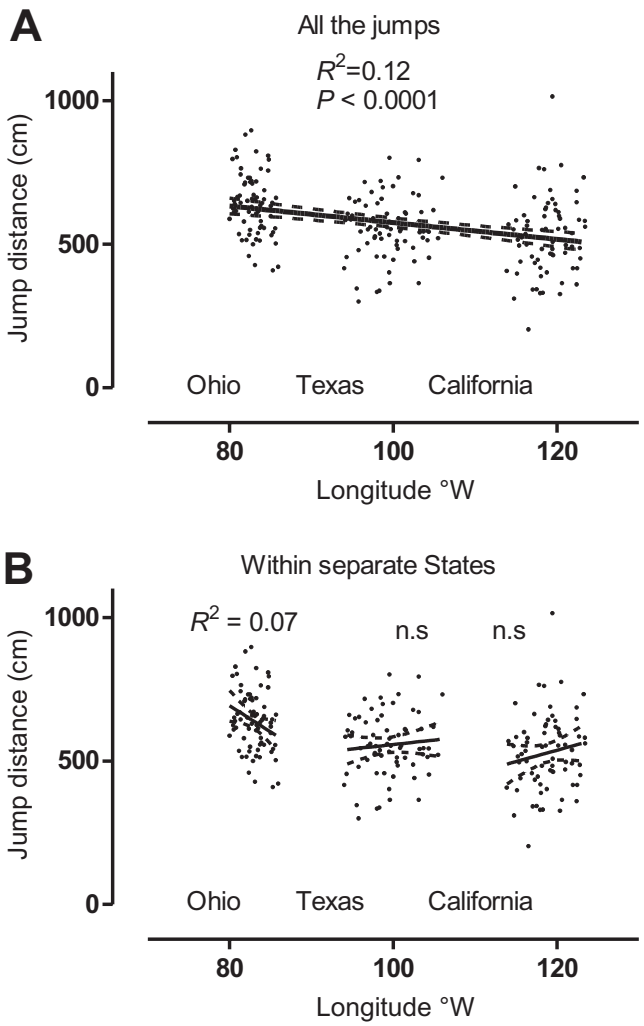

Figure 3

(A) The linear regression relationship, and 95% confidence limits around the line, for jump length related to longitude. (B) If the association is considered within each State, a significant relationship is not found in two States (the range of longitude is insufficient).

of the variation in the dependent values is attributable to the independent value. This approach can be extended by considering several factors that could 'explain' the variation.

In our example, we wish to explore the association between origin and jumping ability. We could express the origin of the frogs in terms of the longitude of where they were found. We choose to use longitude as a convenient continuous variable to 'quantify' origin. (Strictly speaking, this violates one of the assumptions of regression, which is that the independent variable should be normally distributed. In our example we have data from three States, which do not meet this assumption, as can be seen from the data plotted in Figure 1C, showing the jump length of frogs found at different longitudes.) We wish to derive what association there is between jump length and the longitude that the frog comes from, so that we may attribute some of the variation in jump length to this factor. A general, simple, equation can be based on the classic equation for a straight line:

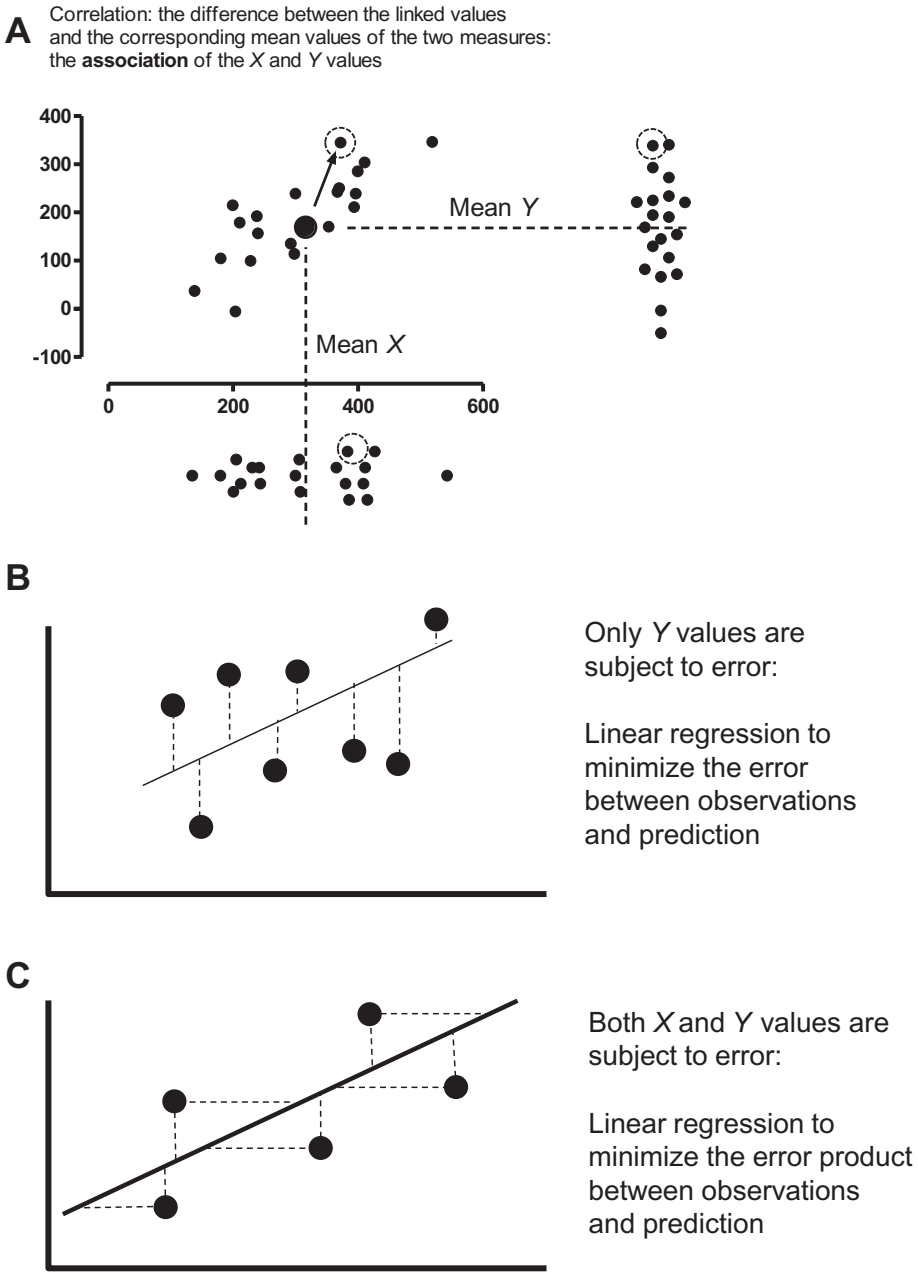

Figure 4

(A) Correlation considers variation in both measures, away from the mean values. (B) The commonly used linear regression procedure does not allow for uncertainty in the independent variable. (C) If both measures have measurement error, then more complex analysis is needed.

$$Y = mX + b$$

If the independent variable is X, and the dependent variable is Y, then

$$\text{Dependent variable} = m(\text{Independent variable}) + b$$

And in our example

$$\text{Jump length} = m(\text{Longitude}) + b$$

The constants m and b are calculated to minimize the differences between the observed jumps and the jumps that would be predicted, at that longitude. Indeed, the equation allows us to *predict*, to some extent, the value of the dependent variable, for a given value of the independent variable. Extrapolation is unwise: in our example, the relationship is only likely to apply between Ohio and California in the USA, not least because there are few frogs to be found in the Pacific Ocean. The mathematical background of regression is described by Curran-Everett (Curran-Everett, 2011).

One approach, often used in preliminary analysis, is to average the jumps from each State (as is done in ANOVA). ANOVA showed us that there was indeed a difference between Ohio and the other States. However, this approach neglects what we consider an important feature of the data, which is the 'variation' in origin, and which is present in the samples for each State. In Figure 3A, we show the calculated linear regression line with the 95% confidence limits of this line.

It is unlikely that we would have obtained these data if there were no difference between the distances that the frogs jumped, in relation to their longitude of origin. However, the R^2 value is small: although 12% of the variation in jump length can be explained by longitude, there remains a lot of variation at each longitude that cannot be attributed to this factor.

One of the reasons that we were able to pick out this small signal (longitude has a small effect) is that we used large samples, and travelled far to collect our frogs. A smaller sample, or a sample which had less variation in longitude, might not have shown this effect. Figure 3B shows that using the same analysis within a single State, even a large one like Texas, fails to detect this effect. Clearly, caution is needed when interpreting such data: is it biologically plausible that longitude is important? Maybe it is the great divide, or genes, or rainfall, or latitude; correlation and regression do not automatically indicate cause and effect. Guidelines in reporting these tests are available (Lang, 2007).

The form of linear regression analysis we have just applied is almost universal, but is not always appropriate. Correlation considers variation in *both* measures by relating the pair of values in each set to their distance from the mean of the measures (Figure 4A). However, linear regression generally only considers variation in the *dependent* variable (plotted on the *Y* axis) and fits a line to minimize the difference between (in our example) jump length observed and jump length predicted (Figure 4B). With global positioning satellites, it may be justifiable to believe that we can estimate longitude exactly enough, but what if we had been using a sextant and a chronometer to measure longitude? We would then be less certain of the accuracy of longitude, and should use analysis in which both *X* and *Y* values can be considered to be variable. Unfortunately, biological signals may vary in both values, particularly if measurements are being compared. In such circumstances, an alternative means of linear regression should be used (Ludbrook, 2010).

References

Archie JP Jr (1981). Mathematic coupling of data. A common source of error. Ann Surg 193: 296–303.

Curran-Everett D (2010). Explorations in statistics: correlation. Adv Physiol Educ 34: 186–191.

Curran-Everett D (2011). Explorations in statistics: regression. Adv Physiol Educ 35: 347–352.

Drummond GB, Vowler SL (2012). Analysis of variance: variably complex. Br J Pharmacol 166: 801–805.

Lang T (2007). Documenting research in scientific articles: guidelines for authors: 3. Reporting multivariate analyses. Chest 131: 628–632.

Ludbrook J (2010). Linear regression analysis for comparing two measurers or methods of measurement: but which regression? Clin Exp Pharmacol Physiol 37: 692–699.

Moses LE, Emerson JD, Hosseini H (1984). Analyzing data from ordered categories. N Engl J Med 311: 442–448.

Walsh TS, Lee A (1998). Mathematical coupling in medical research: lessons from studies of oxygen kinetics. Br J Anaesth 81: 118–120.