Creating and Classifying Measures of Linear Association

Wiebe R. Pestman

Faculty of Science, Utrecht University, The Netherlands

Among the class of most used and abused statistical quantities, there is no doubt that a solid place is reserved for Pearson's correlation coefficient. It therefore seems natural to subject the coefficient to a deliberative analysis as to its basic properties and its place relative to other measures of linear association. A first peculiarity that emerges in such an analysis is that Pearson's coefficient is not a geometric quantity in the sense of modern geometry (see [4]). To be more explicit, the coefficient fails to be invariant when the cloud of points in a scatterplot is rotated (see [1]). It therefore seems odd to use the coefficient as a measure of the degree of clustering around a straight line, this being a geometric phenomenon. Other measures that do better in this respect can easily be created. Such measures, however, fail as a rule to be invariant under the action of rescaling the data. That is to say, they depend on the choice of the units in which the data is expressed. As can be shown by using elementary techniques from mathematical group theory, this is no coincidence. To be more precise, it can be proved that a usable measure of linear association is either a geometric characteristic that is not invariant under scaling or a scaling invariant measure that fails to be geometric. Pearson's coefficient evidently belongs to the latter class. Using basic mathematical group theory it is easy to create various measures of linear association. It is interesting to see that in bivariate cases many such measures are related to the Pearson coefficient. The mentioned techniques, however, are not bound to bivariate statistics; they are in a natural way applicable to the creation and classification of multivariate measures of linear association. As such it is possible to construct, for example, measures to detect multicollinearity among the explanatory variables in linear models. In addition, group theory can also be of use in the development of statistical quantities other than measures of linear association. Thus, extending its triumphal use in various branches of science, Galois group theory can also be successfully applied in statistical sciences (see also [2], [3]). The groups one encounters are usually Lie groups. As will briefly be explained, this can be very helpful when dealing with the computational aspects of one thing and another.

References

[1] Loh WY (1987) Does the Correlation Coefficient Really Measure the Degree of Clustering around a Line? Journal of Educational Statistics, Vol 12, No.3, pp. 235-239.

[2] Pestman WR (1998) Mathematical Statistics, De Gruyter Verlag, Berlin.

[3] Pestman WR (2008) Classification of Bivariate Measures of Linear Association (submitted for publication).

[4] Sénéchal B (1979) Groupes et Géometries, Hermann, Paris.