PRACTICAL USE OF CORRELATION COEFFICIENTS IN THE SOCIAL SCIENCES

Oscar Hernández[1] and Russell Alpizar-Jara[2]
[1]Universidad de Costa Rica, Costa Rica
[2]Universidade de Évora, Portugal
oscar.hernandezrodriguez@ucr.ac.cr

The Pearson correlation coefficient (r) is usually the first measure of association taught at elementary statistics courses. The usual presentation includes scatterplots, computation and interpretation of r, properties, examples, and warnings about inferring causality from high association between two variables. On this last aspect, few introductory textbooks go deeper into the criteria for establishing causation, and there is a lack of convincing examples in the area of the Social Sciences. Although some textbooks give adequate explanations, most of their examples belong to the field of Biostatistics. There is a need to incorporate convincing cases of the practical use of correlation as supporting evidence of causal relationships in the Social Sciences. We contribute with two examples that could be useful for teaching purposes.

INTRODUCTION

An introductory statistics course for students of the Social Sciences should give examples that are really convincing of the practical use of statistical concepts and methods. Nowadays, that the computational aspects are solved by statistical packages like SPSS, or languages like R, more time should be assigned to the understanding of statistical concepts, the logic behind the statistical methods and the illustrations of those concepts and methods with real applications in the various fields of the Social Sciences.

Correlation is one of the most widely misused statistical terms. Many people use it when they really mean the general term *association*, which is a vague term used to describe a relationship between two variables. Most textbooks do not give a general definition of association. Some of them introduce the concept via examples to describe the relationship between two variables, e.g. by showing that taller people tend also to be heavier, or that heavier cars have fewer accidental deaths. It is then added that in cases like these, the pair of variables are said to be *associated*.

As an exception to those textbooks, the book by Moore and McCabe (*Introduction to the Practice of Statistics,* 2006), give a definition of the association between two variables which is more precise:

"Two variables measured on the same individuals are associated if some values of one variable tend to occur more often with some values of the second variable than with other values of the same variable."

They also give a general definition of positive and negative association:

"Two variables are positively associated when above-average values of one tend to accompany above-average values of the other and below-average values also tend to occur together. Two variables are negatively associated when above-average values of one accompany below-values of the other, and vice versa."

The first definition is useful because not only it applies to quantitative variables X and Y, but also when one or both variables are categorical. The second one is mainly for quantitative variables, and it is very useful for interpreting correlations in practical applications.

THE PEARSON CORRELATION COEFFICIENT

Since we have stressed the need to define statistical concepts precisely, and to fully explain the logic behind the definition of statistical indexes or formulas, some words are appropriate for the derivation of the Pearson correlation coefficient.

In most introductory statistics textbooks the study of correlation usually starts with a scatterplot of data of two quantitative variables X and Y. This plot is very useful to appreciate the form, direction and strength of any possible relationship between them. Usually it is made with X and Y measured in the original units. However, we believe that it is also convenient to display a

In M. A. Sorto, A. White, & L. Guyot (Eds.), *Looking back, looking forward. Proceedings of the Tenth International Conference on Teaching Statistics (ICOTS10, July, 2018), Kyoto, Japan.* Voorburg, The Netherlands: International Statistical Institute. iase-web.org [© 2018 ISI/IASE]

scatterplot of the n standardized values $z_{x_i} = \left(\frac{x_i - \bar{x}}{s_X}\right)$ and $z_{y_i} = \left(\frac{y_i - \bar{y}}{s_Y}\right)$, specially for visual comparisons of sets of data, with different units of measurement or order of magnitude.

Some textbooks give the formula for the Pearson correlation r as: $r = \frac{1}{n-1}\sum_{i=1}^{n} z_{x_i} z_{y_i}$, but no explanation is given of the logic behind it. The definition by Moore and McCabe (2006) can be used to show that a direct (positive) association implies r > 0, and that an inverse (negative) association implies r < 0. The appropriateness of r for measuring linear association can be easily shown by proving first that: $-1 \leq r \leq 1$ (a simple application of Cauchy-Schwarz Inequality). Secondly, that when the points (x_i, y_i) fall on the straight line $y_i = a + b\, x_i$, b > 0, then r = 1, by substituting in the r formula: $y_i - \bar{y} = b\,(x_i - \bar{x})$, since $\bar{y} = a + b\,\bar{x}$, and $S_y^2 = b^2 S_x^2$. Similarly, for b < 0, r = -1. It can therefore be easily understood that the closer the points to a straight line, the closer r will be to 1 or -1, and that in this sense, r is a measure of linear association.

THE PRACTICAL USE OF THE CORRELATION COEFFICIENT (r)

Concerning the practical use of the correlation r, it is important to stress to the student that r can be useful as:

- A descriptive measure of linear association between two quantitative variables.
- Supporting evidence of a postulated linear relationship between two variables, or of a theoretical causal model as in path analysis.
- An aid for establishing possible causal relations between variables.

Many introductory statistics textbooks for students of the Social Sciences (e.g. anthropology, demography, political science, sociology, etc.) do not offer enlightening illustrations of the above uses. Textbooks for introductory statistics courses in the biomedical sciences are more illustrative.

Concerning the use of r as descriptive measure of linear association between two variables, there are abundant illustrations. For instance, Agresti and Finlay (1999) give a scatterplot of Y = murder rate and X = poverty rate and concentrate on the prediction equation Y = a + b X, and the value of r. The corresponding value of r = 0.635 is interpreted as implying that X and Y are positively related, that a standard deviation increase in X corresponds to a 0.635 increase in Y and that the correlation is moderately strong. With respect to predicting Y with X, since $r^2 = 0.395$, it is said that the linear prediction equation has 39.5% less error that the mean $\bar{Y}$. However, nothing else is said about the practical consequences for the society.

The use of correlations as evidence supporting a linear relationship between two variables, or a theoretical causal model, postulated by theoretical reasoning, should receive more practical illustrations. We suggest applications like the inverse relationship between price and quantity in economics, or simple path analysis examples. The law of demand, for instance, was derived first by theoretical reasoning, and later confirmed by econometricians with data.

On the use of r as an aid for establishing causation between two variables, there is a lack of illustrations for the Social Sciences in the introductory statistics textbooks. These tend to concentrate more on examples showing that association does not imply causation. For instance, the very well-known case relating the population of Oldenburg, Germany, to the number of storks, which gave a spurious correlation, with a value of r = 0.97. Or as in Agresti and Finlay (1999), who give a correlation of 0.81 between X = height and Y = math test score, for a random sample of students from Lake Wobegon school district, suggesting that taller students tend to have higher scores. However, controlling for the effect of age, they showed that correlations between X and Y were close to zero, within each of three grade levels (2, 5 and 8).

The best examples of causation come from the biomedical sciences. For instance, the use of the association found between smoking and lung cancer, or between smoking and heart disease, as a supporting evidence for establishing causation.

In this context we believe that all introductory statistics courses should discuss the various criteria that have been proposed for establishing the existence of a causal relationship, specifically, the recommendations given by the advisory committee to the Surgeon of the Public Health Service (*Smoking and Health*, 1964):

"As already stated, statistical methods cannot establish proof of a causal relationship in an association. The causal significance of an association is a matter of judgment which goes beyond any statement of statistical probability. To judge or evaluate the causal significance of the association between cigarette smoking and lung cancer a number of criteria must be utilized, no one of which by itself is pathognomonic or a sine qua non for judgment."

According to the advisory committee, these criteria should include: a) the consistency of the association, b) its strength, c) its specificity, d) its temporal relationship, and e) its coherence.

The above criteria are mentioned in a few introductory statistics books, notably in Moore and MacCabe (2006), and Morton and Hebel (1984), in the context of smoking and lung cancer.

For the Social Sciences, Agresti and Finlay (1999), have stated the following criteria for variables X and Y, where X is the cause and Y is the effect: a) X and Y must be statistically dependent, b) X and Y must have the appropriate time order, with cause X preceding effect Y, c) elimination of alternative explanations of the association between X and Y (existence of a spurious relationship: dependence of X and Y on a third variable Z, or X and Y conditionally related on values of another variable W, the relationship disappearing at certain values of W; existence of a relationship between X and Y because of the intervention of another variable S between X and Y; X and Y related simply because of sampling error).

Agresti and Finlay (1999), illustrate the above criteria with the association between smoking and cancer, but they do not give any data, or a value of a correlation. Examples where the use of correlation was a contributed element for establishing causality in a Social Sciences setting are not given. Other textbooks lack illustrations of this sort, and this is precisely the concern that we wanted to address in this paper. A possible explanation for this lack of examples may be due to the more frequent use of qualitative and unobservable latent variables, rather than quantitative and directly measured variables, in the Social Sciences.

EXAMPLES

We provide an illustration to aid teaching with respect to the practical use of correlation, when two variables fulfill the above three criteria to establish that X is a cause of Y. The data refers to the Costa Rican National Fecundity and Health Survey 1986 (Asociación Demográfica Costarricense, 1987, available at https://censos.ccp.ucr.ac.cr/), where X = age of women at first union and Y = age of women at first live birth after union. Strong positive correlations were found ranging from 0.89 to 0.97 in seven provinces. Overall country level correlation coefficient was 0.93 with a 95%CI [0.925, 1.0], Figure 1. Solid evidence of causal inference has been long documented (Davis and Blake, 1956; Bongaarts and Potter, 1983), and determinants of first birth intervals have been a primer concern in developing countries, e.g. recent studies in Ethiopia (Gurmu and Etana, 2014) and in Uganda (Mubiru et al., 2016).

Another example is the use of correlation coefficients as evidence to support causal relationships in path analyses, or in structural equation modelling (Bollen, 1989). For instance, Alpizar-Jara et al. (1990) extensively used correlation coefficients in a confirmatory path analysis of a conceptual model about attitude of a population towards married women's work other than housewives' activities. The study was carried out in May of 1989, in the metropolitan region of San José, Costa Rica, among residents between 18 and 54 years old (n = 338). The study revealed a moderate to strong effect of a latent construct, named chauvinism, on attitude towards married women working outside of the household, indicating that the higher the level of the chauvinism scale the more negative the attitude towards the work of married women outside of the household.

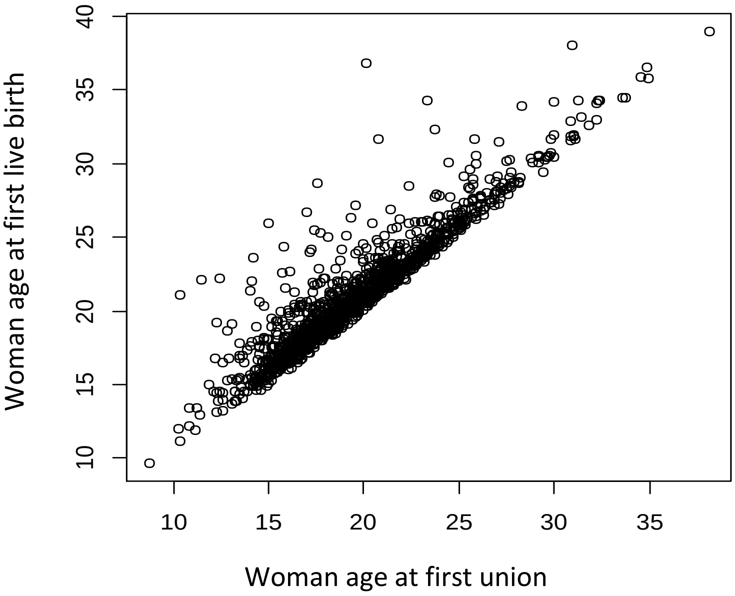

Figure 1. Costa Rican National Fecundity and Health Survey, 1986 (n=1420)

FINAL WORDS

We expect that our paper will motivate introductory statistics teachers to produce more enlightening examples of the use of correlation as supporting evidence of causal relations in the Social Sciences.

ACKNOWLEDGEMENTS

We are grateful to professor Gilbert Brenes, at the University of Costa Rica, for providing the data for our first example, and for several comments that improved the manuscript's presentation.

REFERENCES

Advisory Committee to the Surgeon General of the Public Health Service. (1964). *Smoking and Health*. P.H.S. Publication No.1103, pp.182-189. Public Service, Washington, D.C.

Agresti, A., & Finlay, B. (1999). *Statistical Methods for the Social Sciences*. New Jersey: Prentice Hall, Inc.

Alpizar-Jara, R., Barrantes, M., & Muñoz, B. (1990). Actitudes hacia el trabajo fuera del hogar de la mujer casada: aplicación de un análisis de trayectoria. *II Jornada de Análisis de Datos,* Escuela de Estadística. Universidad de Costa Rica.

Bollen, K.A. (1989). *Structural Equations with Latent Variables*. New York: Wiley.

Bongaarts, J., & R. G. Potter. (1983). *Fertility, Biology, and Behavior: An Analysis of the Proximate Determinants*. New York: Academic Press.

Davis, K., & Blake, J. (1956). Social Structure and Fertility: An Analytic Framework. *Economic Development and Cultural Change, 4*(3), 211-235.

Gurmu, E., & Etana, D. (2014). Age at First Marriage and First Birth Interval in Ethiopia: Analysis of the Roles of Social and Demographic Factors. *African Population Studies, 29*(3), 1332-1344.

Morton, R.F., & Hebel, J.R. (1984). *A Study Guide to Epidemiology and Biostatistics.* University Park Press, Baltimore.

Moore, David, S., & McCabe, G.P. (2006). *Introduction to the Practice of Statistics*. New York: W.H Freeman and Company.

Mubiru, F., Atuhaire, L.K., Lubaale, Y.M., & Wamala, R. (2016). Predictors of time to first birth after marriage among women in Uganda. *African Population Studies, 30*(2) *Supp*., 2482-2494.