Statistical Papers 48, 305-311 (2007)

Statistical Papers

Admissibility Intervals for Linear Correlation Coefficients

Luisa Canal, Rocco Micciolo

Department of Cognitive Sciences and Education
University of Trento
Via Matteo del Ben, 5
38068 – Rovereto (TN) - Italy
(e-mail: luisa.canal@unitn.it)

The correlation coefficient is widely used to quantify the degree of association between two quantitative variables. By resorting to the geometric representation of the linear correlation coefficient, it is possible to calculate the upper and lower bounds of the correlation coefficient between two variables x_1, x_2 when the correlation coefficients with a third variable x_3 are available. Implications in observational studies, where x_3 could be a proxy of a target variable x_2, whose direct measurement is too expensive or impractical, are discussed.

Key words Geometric representation, observational study, proxy variable

1 Introduction

In observational studies, particularly in the epidemiological field, correlations are often used to summarize the results by expressing the relationships between two continuous variables, say x_1 and x_2. For example, x_1 may be a biological variable related to the risk of developing a disease (such as cholesterol or triglyceride plasma levels or blood pressure) and x_2 a variable possibly associated with x_1 (for example, total body fat, total energy intake, personal exposure to a pollutant). However, a direct measurement of x_2 may be too expensive or impractical (for example total body fat could be evaluated using dual energy X-ray absorptiometry, underwater weighing or total body potassium; total energy intake with indirect calorimetry or daily recorded dietary intake). In this case the researcher usually resorts to another variable x_3, that can be considered a *proxy* of the *target* variable x_2 (for example, body mass index or skinfold thickness measured at various body locations; food frequency questionnaires; air quality measures obtained from centrally located outdoor monitoring stations). As a result, the analysis is based on the correlation coefficient r_{13} between x_1 and the proxy x_3, instead of the correlation coefficient r_{12} between x_1 and the target x_2. Obviously, in general, $r_{13} \neq r_{12}$.

It is well known that, for fixed values of r_{13} and r_{23}, r_{12} cannot take any value between –1 and +1, but does have upper and lower bounds depending on r_{13}, r_{23}. In fact, the correlation matrix **R** between x_1, x_2, x_3 must be semi-definite positive (i.e. its eigenvalues must be all non negatives). Therefore, for fixed values of r_{13} and r_{23}, there exist only a limited interval of values *admissible* for r_{12}; it is possible to find this interval from the solutions of the characteristic equation associated with **R**, seeking for what values of r_{12} all the solutions (i.e. the eigenvalues) are non-negative. Even if the solutions of the characteristic equation (which is of third grade) must be all real (since **R** is symmetric), they are somewhat cumbersome to calculate in the general case.

On the contrary, using a geometric approach it is possible to give a simple, general, answer to such a question.

2 A geometric solution

It is well known that the correlation coefficient can be geometrically interpreted as the cosine of an angle; more precisely, the correlation coefficient r_{xy} between the variables x and y is the cosine of the angle α_{xy} between their representative vectors $\mathbf{v}_x$ and $\mathbf{v}_y$: $r_{xy} = \cos(\alpha_{xy})$. This relationship can be reversed; if one knows the value of r_{xy}, it is possible to found the corresponding angle: $\alpha_{xy} = \cos^{-1}(r_{xy})$.

Let $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3$ denote three vectors (which represent the variables x_1, x_2, x_3) separated by the angles $\alpha_{12}, \alpha_{13}, \alpha_{23}$. Let $\mathbf{v}_1$ and $\mathbf{v}_3$ be linearly independent (i.e. $|r_{13}| \neq 1$) and let $\mathbf{v}_2$ be linearly dependent from $\mathbf{v}_1, \mathbf{v}_3$. In this case, as shown by Leung and Lam (1975), the angle α_{12} between $\mathbf{v}_1$ and $\mathbf{v}_2$ can take only two values: $\alpha_{13} - \alpha_{23}$ or $\alpha_{13} + \alpha_{23}$ when $\alpha_{13} + \alpha_{23} \leq 180°$ and $\alpha_{13} - \alpha_{23}$ or $360° - (\alpha_{13} + \alpha_{23})$ when $\alpha_{13} + \alpha_{23} \geq 180°$. If $\mathbf{v}_2$ is not linearly dependent from $\mathbf{v}_1, \mathbf{v}_3$, the angle α_{12} can take all the values between the two defined above. In either case:

$$\cos(\alpha_{13} + \alpha_{23}) \leq r_{12} \leq \cos(\alpha_{13} - \alpha_{23}) \tag{1}$$

From this inequality it is possible to calculate the interval of admissible values for the correlation coefficient r_{12}; we call this interval the *admissibility interval* for r_{12}.

The inequality in equation (1) describes an elliptical region in the (r_{13}, r_{12})–plane if r_{23} is held fixed. In fact, it is possible to rewrite equation (1) as

$$r_{13}r_{23} - \sqrt{(1 - r_{13}^2)(1 - r_{23}^2)} \leq r_{12} \leq r_{13}r_{23} + \sqrt{(1 - r_{13}^2)(1 - r_{23}^2)} \tag{2}$$

The boundary of this region is given by $r_{12}^2 - 2r_{23}r_{12}r_{13} + r_{13}^2 = 1 - r_{23}^2$. The length of the semi-major axis is $\sqrt{1 + |r_{23}|}$ and points in the direction $(\text{sgn}(r_{23}), +1)$, while the length

of the semi-minor axis is $\sqrt{1-|r_{23}|}$ and points in the direction $(\text{sgn}(r_{23}),-1)$.

3 An illustration

Let us consider the case $r_{23} = .95$ (i.e. an high positive correlation between the target variable x_2 and its proxy x_3) and $r_{13} = .70$. Inequality (2) becomes

$$0.665 - \sqrt{0.51 \times 0.0975} \leq r_{12} \leq 0.665 + \sqrt{0.51 \times 0.0975}$$

and the admissibility interval for the linear correlation coefficient r_{12} between the variables x_1 and x_2 is $(.442, .888)$ in the case considered.

This example shows that, even in presence of an "high" correlation coefficient r_{23} between the target variable x_2 and its proxy x_3, the true correlation r_{12} between x_1 and x_2 could be considerably lower than the value r_{13} observed between x_1 and x_3.

4 Ellipses of admissibility

Using inequality (2), it is possible to calculate, for fixed values of r_{23}, the admissibility interval for r_{12} with respect to r_{13} and to plot them. Figure 1 shows the results obtained for four selected values of r_{23} (.8, .9, .95, .99). For a given value of r_{13} on the abscissa, it is possible to read on the ordinate the upper and the lower bounds within which r_{12} must be contained to satisfy the positive-definite constraint.

As the value of $|r_{23}|$ increases, the length of the minor axis of the ellipse decreases; as $|r_{23}|$ approaches one, the ellipse degenerates into the line segment from [–1,–1] to [1,1].

Inequality (2) implies a note of caution in interpreting the correlation r_{13}. In fact, as a consequence of this inequality, it is possible that, while the correlation r_{13} found between x_1 and x_3 may be, for example, positive, the correlation coefficient r_{12} between x_1 and x_2 may be zero (or even negative).

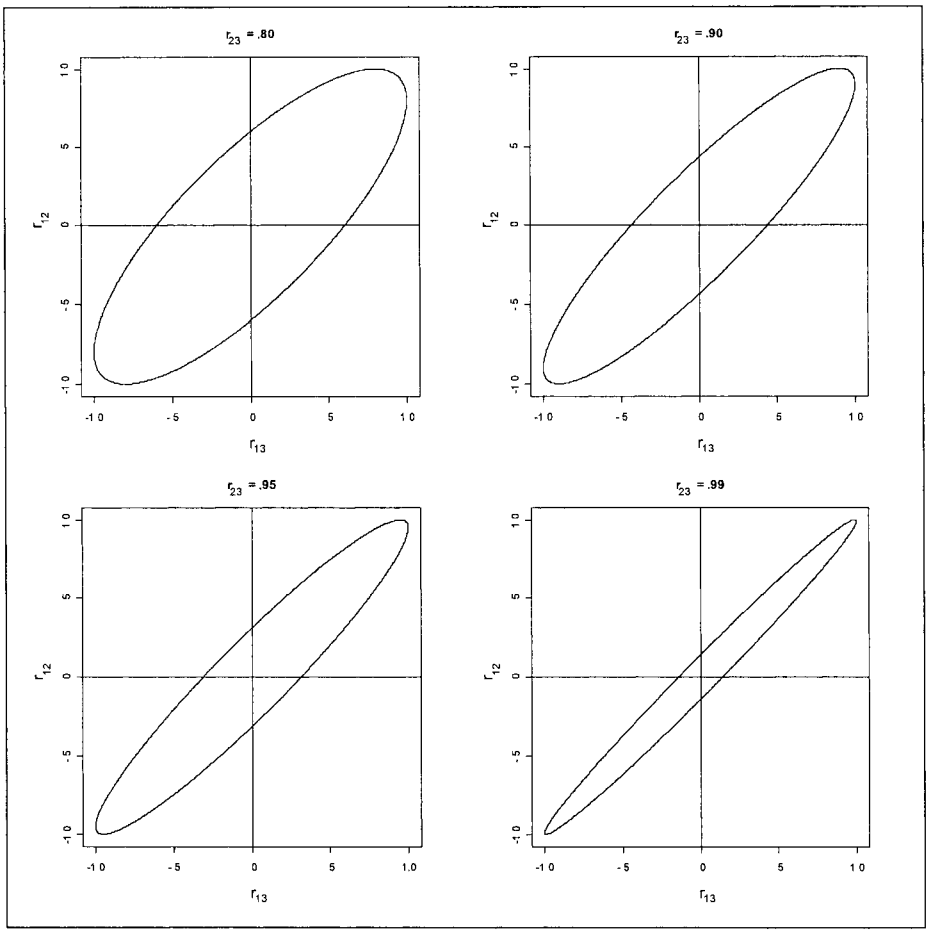

Figure 1. Admissible values for r_{12} with respect to r_{13} for selected values of r_{23}.

This is graphically exemplified in figure 2, where the shaded area shows the range of r_{13} values for which the admissibility interval for r_{12} includes zero (when $r_{23} = .9$); in this case this happens in the range $(-.436,+.436)$ of r_{13}.

For a given value of the correlation coefficient r_{23} between x_2 and its proxy x_3, it is possible to deduce algebraically this range from (2). Let $k = \sqrt{1 - r_{23}^2}$; then if $-k \leq r_{13} \leq +k$, the admissibility interval (1) for r_{12} includes zero.

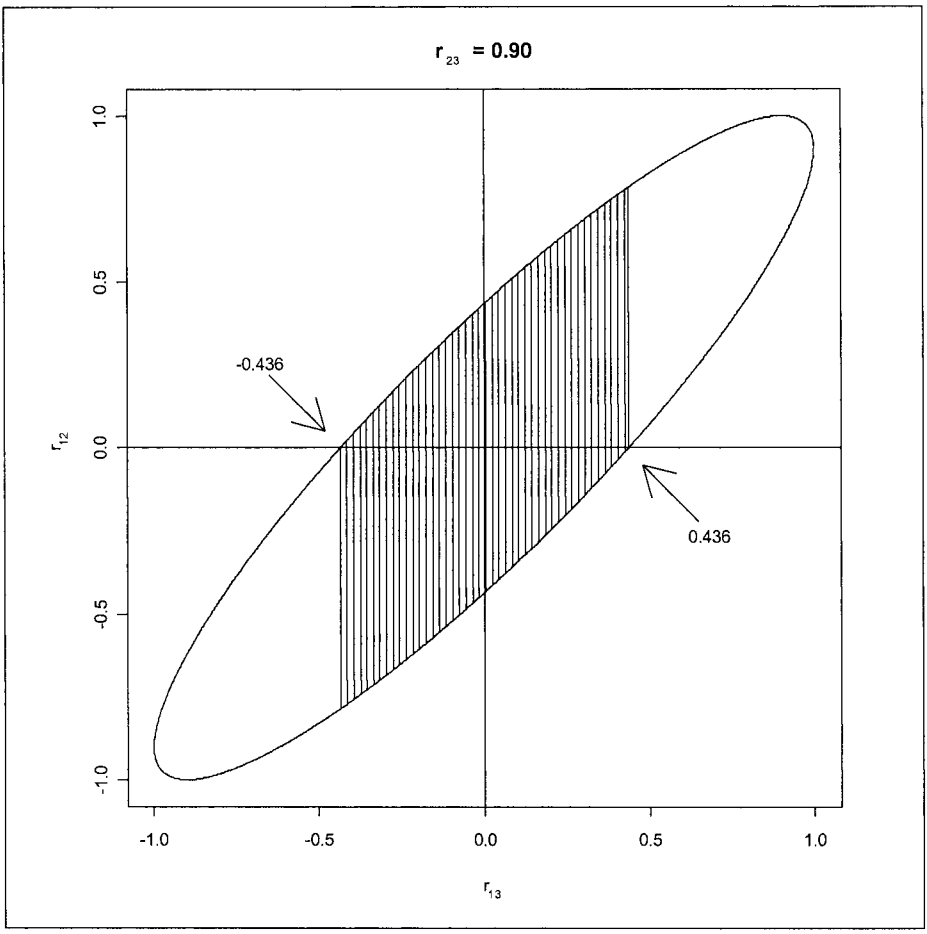

Figure 2. Admissible values for r_{12} with respect to r_{13} when $r_{23} = .9$. The dashed region shows r_{13} values for which the admissibility interval for r_{12} includes zero.

If $r_{23} = .95$ this occurs when r_{13} is between –.312 and +.312, while if $r_{23} = .99$ this occurs when r_{13} is between –.141 and +.141. If the correlation r_{23} is .8, observed values of r_{13} between –.6 and +.6 are consistent with a correlation coefficient r_{12} equal to zero or even with an opposite sign with respect to that found for r_{13}.

The fact that r_{12} and r_{13} may have opposite signs is reminiscent of situation where the line of regression may be opposite to that of the correlation coefficient when there are correlated errors in both variables (McCartin, 2005).

5 Concluding remarks

Although the results presented here are based on fundamental relationships who have been known for many years, their practical implications, particularly in the epidemiological field, are not routinely included in the interpretation of the results. It is possible that both the variability and the difficulties in reproducing the results of epidemiological studies could be, at least in part, due to the difference between the observed correlation with a variable which is a proxy of another target variable and the true correlation with such a target variable.

These results give a quantitative meaning to the prescription of using a proxy highly correlated with the target variable. On the other hand, they indicate that some caution is needed in interpreting low but significant correlations found in empirical studies.

The results presented here are intended to be descriptive and do not give a direct answer to the evaluation of the correlation with the target variable when only a proxy is available. However, by resorting to a geometric representation, an indirect answer is possible. The formulas presented are easily applicable in practice and can be helpful in the critical analysis and in the comparison of the results of empirical studies.

Acknowledgment

The authors are grateful to an anonymous referee for his useful suggestions.

References

Leung, C.K., Lam, K. (1975) A note on the geometric representation of the correlation coefficients, *The American Statistician*, 29, 128–130.

McCartin, B.J. (2005) The geometry of linear regression with correlated errors, *Statistics*, 39, 1–11.