IMS Collections
Nonparametrics and Robustness in Modern Statistical Inference and Time Series Analysis: A Festschrift in honor of Professor Jana Jurečková
Vol. 7 (2010) 62–69
© Institute of Mathematical Statistics, 2010
DOI: 10.1214/10-IMSCOLL706

Measuring directional dependency

Yadolah Dodge[1] and Iraj Yadegari[2]

Abstract: In this article we propose new methods for finding the direction of dependency between two random variables which are related by a linear function.

1. Introduction

The concepts of regression and correlation has been discovered by Francis Galton and Karl Pearson at the turn of the 20th century. The Galton–Pearson correlation coefficient is probably the most frequently used statistical tool in applied sciences, and up to now different interpretations for it has been provided. Rodgers and Nicewander [8] provided thirteen interpretations for it. Rovine and von Eye [9], and Falk and Well [5] show a collection of algebraic and geometric interpretation of the correlation coefficient. An elegant property of the correlation coefficient similar to that of given random variable which is defined by its mean and variance can be found in Nelsen [7] who shows shows that the correlation coefficient is equal to the ratio of a difference and a sum of two moments of inertia about certain line in the plane. Dodge and Rousson [1] provided four new asymmetric interpretations in case of symmetrical error in the linear relationship of two variables including the cube of the correlation coefficient. Using the relationship found in their paper, and assuming the existence of linear relation between two random variables, they determined the direction of dependence in linear regression model. That is, they provided a model on the basis of which one can make a distinction between dependent and independent variables in a linear regression. The directional dependence between two variables, when they follow the Laplace distributions, were provided by Dodge and Whittaker [3] using graphical model approach. Muddapur [6] arrives at the same relationship and found yet another formula between the correlation coefficient and the ratio of two coefficients of kurtosis. However, the author does not indicate how it could be used in determining the direction of dependence between two variables in simple linear regression.

Dodge and Yadegari [4] presented five new asymmetric faces of the correlation coefficient. One of these formulas is the fourth power of the correlation coefficients and ratio of coefficients of excess kurtosis of response and explanatory variable. Also, they showed that, in the regression through the origin, the coefficient of correlation is equal to the ratio of coefficients of variation of explanatory variable to response variable. Thus, the coefficient of variation of response variable is larger that the coefficient of variation of explanatory variable.

[1] Instiute Of Statistics, University of Neuchâtel, Switzerland, e-mail: yadolah.dodge@unine.ch
[2] Islamic Azad University, Kermanshah, Iran, e-mail: i.yadegari@gmail.com

AMS 2000 subject classifications: Primary 62J05; secondary 62M10.

Keywords and phrases: asymptotic interpretation of the correlation coefficient, causality, correlation coefficient, Kurtosis coefficient, linear regression, response variable coefficient of variation, skewness coefficient.

In Section 2 we review some asymmetric formulas for the correlation coefficient, and in Section 3 the concept of the directional dependency between two variables is presented and procedures for determining the direction of dependency between response and explanatory variables in linear regression are discussed. In Section 4 we provide asymmetric measures of directional dependency in linear regression.

2. Some asymmetric faces of the correlation coefficient

Rodgers and Nicewander [8], Rovine and von Eye [9], Falk and Well [5] and Nelsen [7] provided different faces of the correlation coefficient which was discussed by Dodge and Rousson [1, 2] and Dodge and Yadegar in [4]. Also, we present a new face of correlation coefficient. Later we use some of these formulas for determining the direction of dependency between two variables.

Let us consider two random variables X and Y that are related by

$$Y = \alpha + \beta X + \varepsilon, \tag{2.1}$$

where the skewness and the excess kurtosis coefficients of the random variables X and Y are not zero, α is the intercept, β is the slope parameter and ε is an error variable that is independent of X and has normal distribution with zero mean and fixed variance. The correlation coefficient between two random variables X and Y is defined as follows

$$\rho = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y}, \tag{2.2}$$

where $\mathrm{Cov}(X,Y)$ is covariance between X and Y, σ_X^2 and σ_Y^2 are variances of X and Y, respectively. Under the linear model (2.1) we have

$$\rho = \beta \frac{\sigma_X}{\sigma_Y}. \tag{2.3}$$

Since X is independent of ε, starting with (2.1) we can write

$$\sigma_Y^2 = \beta^2 \sigma_X^2 + \sigma_\varepsilon^2$$

and using (2.3) we have

$$1 - \rho^2 = \left(\frac{\sigma_\varepsilon}{\sigma_Y}\right)^2. \tag{2.4}$$

Afterwards we easily obtain

$$\left(\frac{Y-\mu_Y}{\sigma_Y}\right) = \rho \left(\frac{X-\mu_X}{\sigma_X}\right) + (1-\rho^2)^{\frac{1}{2}} \left(\frac{\varepsilon - \mu_\varepsilon}{\sigma_\varepsilon}\right). \tag{2.5}$$

2.1. Cube of the correlation coefficient

The classical notion of skewness is given in the univariate case by the standardized third central moment. The coefficient of skewness of X is

$$\gamma_X = E\Big(\frac{X-\mu_X}{\sigma_X}\Big)^3. \tag{2.6}$$

Dodge and Rousson [1] have proved that under the assumption of symmetry of the error variable and under model (2.1), the cube of the correlation coefficient is equal to the ratio of the skewness of the response variable and the skewness of the explanatory variable. We can derive it in the same way. From third power of both sides of (2.5) and under expectation we have

$$\gamma_Y = \rho^3\gamma_X + (1-\rho^2)^{\frac{3}{2}}\gamma_\varepsilon,$$

where γ_ε is the skewness coefficient of the error variable. If the error variable is symmetric, $\gamma_\varepsilon = 0$, then

$$\rho^3 = \frac{\gamma_Y}{\gamma_X} \tag{2.7}$$

as long as $\gamma_X \neq 0$.

2.2. The 4th power of the correlation coefficient

The coefficient of excess kurtosis of random variable X is defined by

$$\kappa_X = E\Big(\frac{X-\mu_X}{\sigma_X}\Big)^4 - 3. \tag{2.8}$$

Dodge and Yadegari [4] showed that under the assumption of symmetry of the error variable and under model (2.1), the 4th power of the correlation coefficient is equal to the ratio of the kurtosis of the response variables and the kurtosis of the explanatory variable. From the 4th power of both sides of (2.5) and under expectation, and after simplification and using (2.4) we have

$$\kappa_Y = \rho^4\kappa_X + (1-\rho^2)^2\kappa_\varepsilon.$$

If $\kappa_\varepsilon = 0$, we have (as long as $\kappa_X \neq 0$)

$$\rho^4 = \frac{\kappa_Y}{\kappa_X}. \tag{2.9}$$

This formula has a natural interpretation: add a symmetric error to an explanatory variable and you get a response variable with less kurtosis. Also, the fourth power of the correlation may be described as the percentage of kurtosis which is preserved by a linear model.

2.3. The 5th power of the correlation coefficient

If we assume that X and Y are asymmetric, from the fifth power of both sides of (2.5) and under expectation we can obtain

$$\begin{aligned} E\left(\frac{Y-\mu_Y}{\sigma_Y}\right)^5 &= \rho^5 E\left(\frac{X-\mu_X}{\sigma_X}\right)^5 + C_3^5\left(\rho^3\gamma_X(1-\rho^2) + \rho^2(1-\rho^2)^{\frac{3}{2}}\gamma_\varepsilon\right) \\ &\quad + (1-\rho^2)^{\frac{5}{2}} E\left(\frac{\varepsilon-\mu_\varepsilon}{\sigma_\varepsilon}\right)^5, \end{aligned} \tag{2.10}$$

where $C_n^m = \frac{m!}{n!(m-n)!}$. If we assume that $E\left(\frac{\varepsilon-\mu_\varepsilon}{\sigma_\varepsilon}\right)^3 = E\left(\frac{\varepsilon-\mu_\varepsilon}{\sigma_\varepsilon}\right)^5 = 0$, then from (2.7) and (2.10) we have

$$\left(E\left(\frac{Y-\mu_Y}{\sigma_Y}\right)^5 - C_3^5\gamma_Y\right) = \rho^5\left(E\left(\frac{X-\mu_X}{\sigma_X}\right)^5 - C_3^5\gamma_X\right). \tag{2.11}$$

Hence, we obtain a new expression for the correlation coefficient:

$$\rho^5 = \frac{E\left(\frac{Y-\mu_Y}{\sigma_Y}\right)^5 - C_3^5\gamma_Y}{E\left(\frac{X-\mu_X}{\sigma_X}\right)^5 - C_3^5\gamma_X}. \tag{2.12}$$

This formula represents another asymmetric face of the correlation coefficient.

2.4. The ratio of excess kurtosis to skewness

By dividing equation (2.9) to equation (2.7) we obtain

$$\rho = \frac{\kappa_Y/\gamma_Y}{\kappa_X/\gamma_X}. \tag{2.13}$$

The equation (2.12) signifies that we can express the correlation coefficient as a ratio of a function of Y to the same function of X. This ratio is an asymmetric function of the excess kurtosis and the skewness coefficients of dependent and independent random variables.

2.5. Asymmetric function of Joint Distribution

Another asymptotic formula for ρ under model (2.1) may be obtained by introducing higher order correlations

$$\rho_{ij}(X,Y) = E\left[\left(\frac{X-\mu_X}{\sigma_X}\right)^i \left(\frac{Y-\mu_Y}{\sigma_Y}\right)^j\right].$$

We can obtain a beautiful formula for ρ as

$$\rho = \frac{\rho_{12}(X,Y)}{\rho_{21}(X,Y)}. \tag{2.14}$$

Result (2.14) shows a different asymmetric face of correlation which comes from joint distribution of X and Y (Dodge and Rousson [1, 2]).

2.6. The ratio of two coefficients of variation

The coefficient of variation of random variable X, denoted by CV_X, is defined as

$$CV_X = \frac{\sigma_X}{\mu_X}. \tag{2.15}$$

The correlation coefficient can also be expressed as the ratio of two coefficients of variation of random variables related by a linear regression forced from origin (Dodge and Yadegari [4]). Let us consider two random variables X and Y that are related by regression model

$$Y = \beta X + \varepsilon, \tag{2.16}$$

where ε is an error variable with zero mean and fixed variance that is independent of X and $\beta \in R$ is a constant. In the model (2.16) we have $\mu_Y = \beta\mu_X$, then

$$\rho = \frac{CV_X}{CV_Y}. \tag{2.17}$$

From equation (2.17) we conclude that the coefficient of variation of the response variable will always be greater than the coefficient of variation of the explanatory variable.

3. Determining direction of dependence

Consider the situation that a linear relationship exists between two random variables X and Y in the following form

$$Y = \alpha + \beta X + \varepsilon. \tag{3.1}$$

In (3.1) the random variable Y is a linear function of the random variable X, and X is assume to be independent of the error variable ε. In this situation we say that the response variable Y depends on the variable X, and the direction of dependency is from X to Y. Equation (3.1) can also be thought as a causal relationship between explanatory variable (cause) and response variable (effect). If X causes Y, then we select the model (3.1). On the other hand, if Y causes X, then we select the model

$$X = \alpha' + \beta' Y + \varepsilon'. \tag{3.2}$$

In (3.2) the error variable ε' is independent of the explanatory variable Y. In both models (3.1) and (3.2) we assume that the error variable has a normal distribution with zero mean and fixed variance.

If we wish to investigate the direction of dependency, we may hesitate between model (3.1) and model (3.2). To answer such a question, Dodge and Rousson [1] and Dodge and Yadegari [4] proposed some methods for determining the direction of dependency in the linear regression based on the assumption that the skewness or kurtosis coefficient of the error variable is zero.

In what follows, we change the problem of determining the direction of dependence to the problem of comparing two dependent variances or two dependent coefficients of skewness, kurtosis and variation.

3.1. Using joint distribution

Dodge and Rousson [2] has showed an asymmetric face of correlation coefficient, that no assumption is needed about the error variable (except its independence with the explanatory variable).

$$\rho_{XY} = \frac{\rho_{12}(X,Y)}{\rho_{21}(X,Y)}. \tag{3.3}$$

This formula can be obtained from joint distribution. They used formula (3.3) to determine the direction of dependence between X and Y. Since $|\rho_{XY}| \leq 1$,

$$\rho_{12}^2(X,Y) \leq \rho_{21}^2(X,Y). \tag{3.4}$$

Thus, Y is a response variable. A similar argument can be provided for the linear regression dependence of X on Y. Then, $\rho_{12}^2(X,Y) \leq \rho_{21}^2(X,Y)$ implies Y is the response variable and $\rho_{12}^2(X,Y) \geq \rho_{21}^2(X,Y)$ implies X is the response variable.

3.2. Comparing skewness coefficients

Dodge and Rousson [2] showed that under assumption of symmetry of the error variable and under model (3.1), the cube of the correlation coefficients is equal to the ratio of the skewness of the response variable and the skewness of the explanatory variable:

$$\rho_{XY}^3 = \frac{\gamma_Y}{\gamma_X}, \tag{3.5}$$

(as long as $\gamma_X \neq 0$). They used formula (3.5) to determine the direction of dependence between X and Y. Since $|\rho_{XY}| \leq 1$,

$$\gamma_Y^2 \leq \gamma_X^2. \tag{3.6}$$

Thus, the direction of dependence is from X to Y (Y is a response variable). A similar argument can be provided for the linear regression dependence of X on Y. Then, $\gamma_X^2 \geq \gamma_Y^2$ implies Y is the response variable and $\gamma_X^2 \leq \gamma_Y^2$ implies X is the response variable.

3.3. Comparing kurtosis coefficients

Dodge and Yadegari [4] gave another method that works in symmetric and asymmetric situations. Under model (2.1), the fourth power of the correlation coefficient is equal to the ratio of kurtosis of the response variable to the kurtosis of the explanatory variable, (as long as $\kappa_X \neq 0$)

$$\rho^4 = \frac{\kappa_Y}{\kappa_X}, \tag{3.7}$$

where κ_X and κ_Y are kurtosis coefficients of X and Y respectively (as long as $\kappa_X \neq 0$). Since $\rho^4 \leq 1$

$$\kappa_Y \leq \kappa_X. \tag{3.8}$$

This shows that the kurtosis of the response variable is always smaller than the kurtosis of the explanatory variable. Then, for a given ρ_{XY}, $\kappa_X \geq \kappa_Y$ implies Y is the response variable and $\kappa_X \leq \kappa_Y$ implies X is the response variable.

We can similarly use inequalities (2.13) and the 5th power of the correlation coefficient (2.12) to assessing direction of dependence in a linear regression.

3.4. Comparing coefficients of variation

Now consider the situation that a linear relationship exists between two random variables X and Y in the following form

$$Y = \beta X + \varepsilon. \tag{3.9}$$

If X causes Y, then we select the model (3.9). In the other hand, if Y causes X, then we select the model

$$X = \beta' Y + \varepsilon'. \tag{3.10}$$

In (3.10) the error variable ε' is independent of the explanatory variable Y. In both models (3.9) and (3.10) we assume that the error variable has a zero mean and fixed variance. Under assumptions of the model (3.9) and from (3.10), we can conclude that

$$\rho = \frac{CV_X}{CV_Y}. \tag{3.11}$$

Thus, the coefficient of variation of response variable is larger than the coefficient of variation of explanatory variable.

3.4.1. Special case (comparing variables)

Let us consider two random variables X and Y, where a linear relationship exists between them in the following form

$$Y = X + \varepsilon \tag{3.12}$$

or

$$X = Y + \varepsilon'. \tag{3.13}$$

Under model (3.12) we have $\rho^2 = \frac{\sigma^{\hat{2}}_X}{\sigma^2_Y}$ (obtained from (2.3) when $\beta = 1$) and then $\sigma^2_Y > \sigma^2_X$, and under model (3.13) we can obtain that $\sigma^2_Y < \sigma^2_X$. Then, the variance of the explanatory variable is always smaller than the variance of the response variable. Then, $\sigma^2_Y > \sigma^2_X$ implies Y is the response variable and $\sigma^2_Y < \sigma^2_X$ implies X is the response variable.

4. Measures of the directional dependency

We say that the direction of dependency is from X to Y, denoted by $X \to Y$, if a linear relationship exists between random variables X and Y in the following form

$$Y = \alpha + \beta X + \varepsilon, \tag{4.1}$$

where α is the intercept and β is the slope parameter and ε is an error variable that is independent of X and has a normal distribution with zero mean and a fixed variance. For measuring amount of asymmetric dependency between X and Y we cannot use the Galton–Pearson correlation coefficient, because the Galton–Pearson correlation is a symmetric measure of dependency between two random variables. In situations where we have asymmetric measures of dependency, we can present new procedures for determining the direction of dependency. Using the skewness and kurtosis coefficients, in this section, we propose two new asymmetric measures of dependency to distinguish the response from explanatory variable.

Let us consider two random variables X and Y that are related by a linear relationship (4.1). We define another directional correlation coefficient as

$$S(X \to Y) = \frac{\gamma^2_X}{\gamma^2_X + \gamma^2_Y}. \tag{4.2}$$

Here are some properties of this measure:

1. $0 < S(X \to Y) < 1$
2. $S(Y \to X) = 1 - S(X \to Y)$
3. If $\gamma^2_Y \leq \gamma^2_X$, then $S(Y \to X) \leq S(X \to Y)$
4. If $\gamma^2_X = \gamma^2_Y$, then $S(X \to Y) = S(Y \to Y) = \frac{1}{2}$
5. If $\gamma^2_Y < \gamma^2_X$, then $\frac{1}{2} < S(X \to Y) < 1$
6. If $\gamma^2_Y > \gamma^2_X$, then $0 < S(X \to Y) < \frac{1}{2}$.

Thus, $S(X \to Y) > S(Y \to X)$ implies Y is the response variable and $S(X \to Y) < S(Y \to X)$ implies X is the response variable.

We can use the kurtosis coefficients to introduce another asymmetric measures of dependency between two random variables, which measures the directional dependency. Under the model (4.1), we define a measure of the directional dependence in this model as

$$K(X \to Y) = \frac{\kappa_X^2}{\kappa_X^2 + \kappa_Y^2}. \tag{4.3}$$

Here are some properties of the kurtosis-based directional correlation:

1. $0 < K(X \to Y) < 1$
2. $K(Y \to X) = 1 - K(X \to Y)$
3. If $\kappa_X = \kappa_Y$, then $K(X \to Y) = K(Y \to X) = \frac{1}{2}$
4. If $\kappa_Y^2 < \kappa_X^2$, then $\frac{1}{2} < K(X \to Y) \leq 1$
5. If $\kappa_Y^2 \leq \kappa_X^2$, then $K(Y \to X) \leq K(X \to Y)$
6. If $\kappa_Y^2 > \kappa_X^2$, then $0 \leq K(X \to Y) < \frac{1}{2}$.

Thus, $K(X \to Y) > K(Y \to X)$ implies Y is the response variable and $K(X \to Y) < K(Y \to X)$ implies X is the response variable.

References

[1] Dodge, Y. and Rousson, V. (2000). Direction dependence in a regression line. *Commun. Stat. Theory Methods* **29** 9–10 1957–1972.

[2] Dodge, Y. and Rousson, V. (2001). On asymmetric property of the correlation coefficient in the regression line. *Am. Stat.* **55** 1 51–54.

[3] Dodge, Y. and Wittaker, J. (2000). The information for the direction of dependence in L1 regression. *Commun. Stat. Theory Methods* **29** 9–10 1945–1955.

[4] Dodge, Y. and Yadegari, I. (2009). On direction of dependence. *Metrika* **72** 139–150.

[5] Falk, R. and Well, A. D. (1997). Faces of the correlation coefficient. *J. Statistics Education* [Online] **5** 3.

[6] Muddapur, M. (2003). Dependence in a regression line. *Commun. Stat. Theory Methods* **32** 10 2053–2057.

[7] Nelsen, R. B. (1998). Regression lines, and moments of inertia. *Amer. Statistician* **52** 4 343–345.

[8] Rodgers, J. L. and Nicewander, W. A. (1988). Thirteen ways to look at the correlation coefficient. *Am. Stat.* **42** 59–66.

[9] Rovine, M. J. and Eye, A. (1997). A 14th way to look at a correlation coefficient: Correlation as the proportion of matches. *Am. Stat.* **51** 42–46.