3.1 Correlation as a Measure of Association

Ming Lu

2022-04-02

# 3.1.1 Linear Correlation

Definition 3.1 (Pearson’s ) The Pearson’s correlation coefficient of two random variables and is defined as

Formally, is a measure of the strength of linear dependence between and . To see this, it is best to look at the following equivalent form of :

* It is easy to estimate from a sample of by combining the formulas for sample variance and covariance, to get the sample linear correlation
* This is an unbiased estimator of in the sense stated in Remark .
* It is invariant under changes of scale and location in and :
* where is the sign function.
* It gives an elegant expression for the variance of a linear combination of random variables (see Problem 2 under Sect. 3.5). We can now rewrite the covariance of and as
* and this shows how the possible interdependency of and is affected by their individual standard deviations. Thus, while covariance signals a possible dependence of one variable to the other, it is the correlation that determines the strength of this dependence.

Autocorrelation function (ACF). When dealing with time series of returns of financial assets, for which we assume are weakly stationary, it is of interest to assess the linear dependence of the series with its past; i.e., we look for autocorrelation. For a return time series , the lag-k autocorrelation of is the correlation coefficient between and , and is usually denoted by ,

Note that the hypothesis of weakly stationarity is needed in order to have and hence, is only dependent on and not of , which also guarantees second moments are finite and well-defined. The function is referred to as the ACF of .

For a given sample of returns and , the sample lag- autocorrelation of is

where is the sample mean of . Testing autocorrelation. There are various statistical tests to check whether the data are serially correlated. One such test is the Ljung-Box test, a so-called portmanteau test in the sense that takes into account all sample autocorrelations up to a given lag . For a sample of returns, the statistic is given by

as a test for the null hypothesis against the alternative for some is rejected if is greater than the th percentile of a chi-squared distribution with degrees of freedom. Alternatively, is rejected if the -value of is , the significance level. For time periods a good upper bound for the choice of lag is . Further details on this and other tests of autocorrelation can be learned from (Tsay 2010). In the function computes and plots estimates of the autocorrelation (by default) or autocovariance of a matrix of time series or one time series object , up to lag . This bound is justified by the Ljung-Box test and it is the default value; to give a specific number of lags use the parameter lag.max. The Ljung-Box test is implemented in by the function Box. test ().

Deficiencies of linear correlation as dependency measure. A serious deficiency of the linear correlation is that it is not invariant under all transformations of random variables and for which order of magnitude is preserve. That is, for two real-valued random variables and arbitrary transformations , we have in general

For example, for normally-distributed vectors and the standard normal distribution function , we have

(see Joag-dev (1984, 1 ). In general, under the normal distribution,

for arbitrary real-valued transformations (cf. Embrechts et al. 2002). The quality of been invariant under arbitrary order-preserving transformations is desirable in order to obtain a distribution-free measure of dependency (and hence, nonparametric), since inferences can be made by relative magnitudes as opposed to absolute magnitudes of the variables. An additional deficiency of linear correlations is that the variances of and must be finite for the correlation to be defined. This is a problem when working with heavy-tailed distributions, which is often the case for financial time series.

# 3.1.2 Properties of a Dependence Measure

Thus, what can be expected from a good dependency measure? Following the consensus of various researchers, summarized in Joag-dev (1984), Embrechts et al. (2002) and Gibbons and Chakraborti (2003, Chap. 11), among others, a “good” dependency measure , which assigns a real number to any pair of real-valued random variables and , should satisfy the following properties: (1) (symmetry). (2) (normalization). (3) if and only if are co-monotonic; that is, for any two independent pairs of values and of whenever or whenever . This property of pairs is termed perfect concordance of and . (4) if and only if are counter-monotonic; that is, for any two independent pairs of values and of whenever or whenever . In this case, it is said that the pairs and are in perfect discordance. (5) is invariant under all transformations of and for which order of magnitude is preserve. More precisely, for strictly monotonic on the range of ,

The linear correlation satisfies properties (1) and (2) only. A measure of dependency which satisfies all five properties (1)-(5) is Kendall’s . The trick is to based the definition on the probabilities of concordance and discordance of the random variables.

# 3.1.3 Rank Correlation

Definition 3.2 (Kendall’s ) Given random variables and , the Kendall correlation coefficient (also known as rank correlation) between and is defined as

where, for any two independent pairs of values from ,

and are the probabilities of concordance and discordance, respectively. Thus, Kendall’s measure of dependency reflects the agreement in monotonicity between two random variables. To estimate from pairs of sample random values , define

Then if these pairs are concordant; if the pairs are discordant; or 0 if the pairs are neither concordant nor discordant. An unbiased estimation of is given by

The distribution of under the null hypothesis of no association 0 ) is known. Tables for values of for can be found in Gibbons and Chakraborti (2003). Moreover, is asymptotically normal with zero mean and variance . Example 3.3 The function cor can also compute correlation with the sample estimator of Kendall’s tau by setting method= ” kendall “. We use the same data as in R Example to compare with Pearson’s linear correlation.