Brief Report
Open Access
CrossMark

Correlation analysis for different types of variables and relationship between different correlation coefficients

Volume 11 Issue 4 - 2022

Shimin Zheng[1], Yan Cao[2]
[1]Department of Biostatistics, East Tennessee State University, USA
[2]Center for Nursing Research, East Tennessee State University, USA

Correspondence: Shimin Zheng, Department of Biostatistics and Epidemiology, East Tennessee State University, USA, Tel 423-439-4327, Email zhengs@etsu.edu

Received: September 4, 2022 | **Published:** September 20, 2022

Introduction

The purpose of this article is to provide a summary about statistical correlation analysis and relationship between simple, multiple and partial correlation coefficients.

Statistical correlation analysis and regression analysis are related, but different. Correlation analysis quantifies the strength of the linear relationship between two variables or between two sets of variables, most often two continuous variables, or between two sets of continuous variables, whereas regression analysis is used to determine the relationship in the form of an equation between two variables or two sets of variables. Unlike regression analysis, to do correlation analysis, we don't have to distinguish cause and effect, or dependent and independent variables. Most often, the simple correlation coefficient is used. It is also called Pearson product-moment correlation coefficient.[1] It is a measure of the strength and direction of association between two variables measured on at least an interval scale. It can range from -1 to 1. However, maximum (or minimum) values of some simple correlations cannot reach unity (i.e., 1 or -1)

Correlation analysis is not always dealing with one-to-one correlation, i.e., the correlation between two variables. It can be partial correlation (adjusted one-to-one correlation). It can also be one-to-many, or multiple correlation.[2] In statistics, the coefficient of multiple correlation is a measure of how well a given variable can be predicted using a linear function of a set of other variables. It is the correlation between the variable's values and the best predictions that can be computed linearly from the predictive variables.

Relationship between simple and multiple correlation coefficients

The formula to compute the simple correlation coefficient between variables x and y is

$$r=\frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2}\sqrt{\sum(y-\bar{y})^2}}=\frac{n\sum xy-\sum x\sum y}{\sqrt{n\sum x^2-\left(\sum x\right)^2}\sqrt{n\sum y^2-\left(\sum y\right)^2}} \quad (1)$$

The t-statistic $\frac{r}{\sqrt{1-r^2}}\sqrt{n-2}$ (df=n-2) is used to conduct hypothesis test $H_0:\rho=0$ vs $H_a:\rho\neq 0$. The formula to compute the multiple correlation coefficient between y and $x_1,x_2....,x_k$ is

$$r=\sqrt{R^2}=\sqrt{1-\frac{\sum(y-\hat{y})^2}{\sum(y-\bar{y})^2}}=\sqrt{1-\frac{SSE}{SST}}=\sqrt{\frac{SSR}{SST}} \quad (2)$$

The F-statistic $\frac{SSR/k}{SSE/(n-k-1)}=\frac{MSR}{MSE}=\frac{(n-k-1)R^2}{k(1-R^2)}\sim F(k,n-k-1)$

is used to conduct hypothesis test $H_0:\rho^2=0$ vs $H_a:\rho^2\neq 0$.

Multiple correlation coefficient between y and x_1,x_2 can be calculated using simple correlation coefficients

$$r=\sqrt{\frac{r_{yx_1}^2+r_{yx_2}^2-2r_{yx_1}r_{yx_2}r_{x_1x_2}}{1-r_{x_1x_2}^2}}. \quad (3)$$

Generally, the multiple correlation coefficient between y and $x_1,x_2,\cdots,x_k$ can be calculated using simple correlation coefficients.[3,4,5]

$$r=\sqrt{1-\frac{det(R)}{R_{11}}}, \quad (4)$$

where

$$R=\begin{bmatrix}1 & r_{01} & r_{02} & \cdots & r_{0k}\\ r_{01} & 1 & r_{12} & \cdots & r_{1k}\\ r_{02} & r_{12} & 1 & \cdots & r_{2k}\\ \vdots & \vdots & \vdots & \ddots & \vdots\\ r_{0k} & r_{1k} & r_{2k} & \cdots & 1\end{bmatrix}$$

and R_{11} is the cofactor of the $(1,1)$th element of matrix R, $det(R)$ is the determinant of matrix R, r_{0j} is the correlation coefficient between y and x_j, $j=1,2,\ldots,k$, r_{ij} is the correlation coefficient between x_i and x_j, $i,j=1,2,\ldots,k$. Let $r^{-1}{}_{ij}=\left(r^{ij}\right)$, then we have

$$r=\sqrt{1-\frac{1}{r^{00}}} \quad (5)$$

©2022 Zheng et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and build upon your work non-commercially.

Let $(q_{ij})0\leqslant i,j\leqslant k$ be the dispersion matrix of $y,x_1,x_2,\ldots,x_k$ and $(q_{ij})^{-1}=(q^{ij})$, then we have

$$r=\sqrt{1-\frac{1}{q_{00}q^{00}}} \tag{6}$$

Relationship between simple, multiple and partial correlation coefficients

Multiple correlation coefficient can be also calculated using simple and partial correlation coefficients Kendall. [3]

$$1-r^2_{y\cdot x_1x_2\ldots x_k}=(1-r^2_{yx_1})(1-r^2_{yx_2\cdot x_1})(1-r^2_{yx_3\cdot x_1x_2})\ldots(1-r^2_{yx_k\cdot x_1x_2\ldots x_{k-1}}). \tag{7}$$

Formally, the partial correlation between x and y given $z_1,z_2.....,z_n$ is written as $r_{xy.z}$, where z is an n-dimensional vector, $z=\{z_1,z_2.....,z_n\}$. Let N denote the number of observations, then

$$r_{xy\cdot\underline{z}}=\frac{N\sum_{i=1}^{N}e_{x,i}e_{y,i}-\sum_{i=1}^{N}e_{x,i}\sum_{i=1}^{N}e_{y,i}}{\sqrt{N\sum_{i=1}^{N}e^2_{x,i}-(\sum_{i=1}^{N}e_{x,i})^2}\sqrt{N\sum_{i=1}^{N}e^2_{y,i}-(\sum_{i=1}^{N}e_{y,i})^2}}=\frac{N\sum_{i=1}^{N}e_{x,i}e_{y,i}}{\sqrt{N\sum_{i=1}^{N}e^2_{x,i}}\sqrt{N\sum_{i=1}^{N}e^2_{y,i}}} \tag{8}$$

where e_x and e_y are residuals resulting from the linear regression of x with $\underline{z}$ and of y with $\underline{z}$ respectively. Especially, if we have z_1 only, the partial correlation between x and y given z_1 is

$$r_{xy\cdot z_1}=\frac{r_{xy}-r_{xz_1}r_{yz_1}}{\sqrt{(1-r^2_{xz_1})(1-r^2_{yz_1})}} \tag{9}$$

The partial correlation between x and y given z_1 and z_2 is

$$r_{xy\cdot z_1z_2}=\frac{r_{xy\cdot z_1}-r_{xz_2\cdot z_1}r_{yz_2\cdot z_1}}{\sqrt{(1-r^2_{xz_2\cdot z_1})(1-r^2_{yz_2\cdot z_1})}} \tag{10}$$

The formula (10) can be extended to more general case: the partial correlation between x and y given $z_1,z_2,\ldots,z_k$ Kendall.[3] is

$$r_{xy\cdot z_1z_2\ldots z_k}=\frac{r_{xy\cdot z_2z_3\ldots z_k}-r_{xz_1\cdot z_2z_3\ldots z_k}r_{yz_1\cdot z_2z_3\ldots z_k}}{\sqrt{(1-r^2_{xz_1\cdot z_2z_3\ldots z_k})}\sqrt{(1-r^2_{yz_1\cdot z_2z_3\ldots z_k})}} \tag{11}$$

The partial correlation can also be calculated using multiple correlation. For example, the partial correlation between x and y given z_1,z_2 is

$$r_{xy\cdot z_1z_2}=\sqrt{\frac{r^2_{x.yz_1z_2}-r^2_{x.z_1z_2}}{1-r^2_{x.z_1z_2}}} \tag{12}$$

The partial correlation between x and y given z_1, z_2 and z_3 is

$$r_{xy\cdot z_1z_2z_3}=\sqrt{\frac{r^2_{x\cdot yz_1z_2z_3}-r^2_{x\cdot z_1z_2z_3}}{1-r^2_{x\cdot z_1z_2z_3}}} \tag{13}$$

Generally, the partial correlation between x and y given $z_1,z_2.....,z_k$ is

$$r_{xy\cdot z_1z_2\ldots z_k}=\sqrt{\frac{r^2_{x\cdot yz_1z_2\ldots z_k}-r^2_{x\cdot z_1z_2\ldots z_k}}{1-r^2_{x\cdot z_1z_2\ldots z_k}}} \tag{14}$$

Suppose we have z_1 only, the t-statistic $\frac{r_{xy\cdot z_1}\sqrt{n-\upsilon}}{\sqrt{1-r^2_{xy\cdot z_1}}}\sim t(n-\upsilon)$ is used to conduct hypothesis test $H_0:\rho_{xy.z_1}=0$ vs $H_a:\rho_{xy.z_1}\neq 0$, where n is sample size, υ is total number of variables employed in the analysis, here $\upsilon=3$ since we have three variables x,y and z_1.

Canonical correlation analysis

In addition, correlation analysis can be used to determine association between many variables and many variables (many-to-many), the canonical correlation analysis (CCA),[6] which includes deep CCA, sparse CCA, kernel CCA, generalized CCA, regularized CCA, nonlinear CCA. The canonical correlation analysis (CCA) is a standard tool of multivariate statistical analysis for discovery and quantification of associations between two sets of variables.

Polychoric and tetrachoric correlation

Correlation analysis is not always used to determine association between continuous or ordinary variables. It can also be used to determine the association between two categorical variables, or between one continuous variable and another categorical variable. The polychoric correlation is used to measure the association between ordered-category variables with an assumption of an underlying joint continuous distribution.[7,8] A categorical variable is often a rough measurement of an underlying continuous variable. For instance, a dichotomous variable (adult or not) is observed as 'Yes' when age is 18 years or above, and as 'No' if age $<$ 18 years. The underlying variable is age, which is continuous. Hence, it is reasonable to assume that a continuous variable underlies a categorical (dichotomous or polychotomous) observed variable. Therefore, we can conduct the estimation of the polychoric correlation coefficient via Markov chain Monte Carlo methods assuming the underlying distribution is multivariate normal. Especially, the polychoric correlation between two observed binary variables is also known as tetrachoric correlation.[9] Suppose we have a 2×2 table with two binary variables, x and y, then

Tetrachoric correlation = $\cos(\pi/(1+\sqrt{(n_{11}\times n_{22})/n_{12}/n_{21}}))$.

Point biserial correlation and biserial correlation

On the other hand, the point biserial correlation is used to determine an association between one continuous variable and another naturally binary variable.[6] For example, the correlation between gender and salary is called point biserial correlation. The formula for the point biserial correlation coefficient is

$$t_{pb}=\frac{Q_1-Q_0}{s_n}\sqrt{pq} \tag{15}$$

where Q_1 is the mean of the positive or 'Yes' group, defined by the dichotomous variable, Q_0 is the mean of the negative or 'No' group, defined by the same dichotomous variable, s_n is the standard deviation for all, p is the 'Yes' proportion and q is the 'No' proportion.

Biserial correlation is very close to point biserial correlation, but one of associated variables is dichotomous ordinal and has an underlying continuity.[11] For example, depression level can be measured on a continuous scale, such as PHQ-9, the nine-item depression scale of the patient health questionnaire, or the Hamilton rating scale for depression, but can be classified dichotomously as

high/low. The formula for biserial correlation coefficient between a dichotomous ordinal variable (W) and one continuous variable (M) is

$$r_b = [(M_1 - M_0) \times (pq / M)] / \sigma_m \quad (16)$$

where M_0 is mean score of M when $W = 0$, M_1 is the mean score of M when $W = 1$, q is proportion for $W = 0$, p is proportion for $W = 1$, σ_m is population standard deviation, M is the height of the standard normal distribution at z, where $P(z' < z) = q$ & $P(z' > z) = p$.

If point-biserial correlation is known, you can also find biserial correlation with the following formula Sheskin D[12]

$$r_b = (\frac{r_{pb}}{h})\sqrt{p_0(1-p_0)}, \quad (17)$$

where

$$h = \frac{e^{-u^2/2}}{\sqrt{2\pi}} \quad (18)$$

$$Pr[Z \geq u \mid Z \sim N(0,1)] = p_1 \quad (19)$$

We can have a natural extension of the model above if we have more than two ordered rating levels. We can assume that the joint distribution of the quantitative variable and a latent continuous variable underlying the ordinal variable is bivariate normal when we compute a polyserial correlation coefficient (standard error) between a quantitative variable and an ordinal variable. Either the maximum-likelihood (ML) estimator or a quicker 'two-step' approximation can be used. For the ML estimator the estimates of the thresholds and the covariance matrix of the estimates are also available.

Conclusion

In this article we have discussed about Pearson product-moment correlation coefficient, simple, multiple, partial correlation, the relationship among them, the concepts and the formulas to compute each specific coefficient. Also, we have discussed the multivariate canonical correlation between many and many variables. In addition, we have discussed about tetrachoric or polychoric correlation between two observed binary variables or between two ordered-multiple-category variables, as well as the polyserial correlation between a quantitative variable and an ordinal variable, point biserial correlation between one continuous variable and one naturally binary variable, and biserial correlation which is very close to point biserial correlation, but one of associated variables is dichotomous ordinal and has an underlying continuity. To extend the relationship between Pearson product-moment correlation coefficient, simple, multiple, partial correlation to the relationship for other kinds of correlation, such as polychoric, polyserial correlation, can be further study.

Acknowledgments

None.

Conflicts of interest

The authors declare no conflicts of interest.

References

1. Pearson K. Early Statistical Papers. Cambridge: England; 1948.
2. Donna L Mohr, William J Wilson, Rudolf J. Freund. *Statistical Methods*. 4th ed. *Academic Press*;2022.
3. Kendall Maurice G, Francis William. *The Advanced Theory of Statistics*. 4th ed. Charles Griffin & Company Limited;1948.
4. John F Kenney. *A Mathematics of Statistics (part two)*. 2nd ed D. Van Nostrand Company 1939.
5. Rao CR. Linear Statistical Inference and Its Applications by Rao. *John Wiley & Sons*,2nd ed;1965:p266–268
6. Bruce Thompson. Canonical Correlation Analysis. *Sage Publications. 1984*.
7. Drasgow F. Polychoric and polyserial correlations. In: Kotz L, Johnson NL, editors. *Encyclopedia of Statistical Sciences.* New York: Wiley. 1988;7:69–74.
8. Olsson U. Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika*. 1979;44(4):443–460.
9. Pearson K. correlation between a measured character A, and a character B. *Biometrika*. 1901;7:96–105.
10. Brown JD. Understanding research in second language learning: A teacher's guide to statistics and research design. *Cambridge: Cambridge University Press*. 1988.
11. Pearson K. On a new method for determining the correlation between a measured character A, and a character B. *Biometrika*. 1909;7:96–105.
12. Sheskin D. Handbook of Parametric and Non-Parametric Statistical Procedure. 5th edn. Boca Raton, FL: CRC Press; 2011.