Pakistan Journal of Applied Sciences 1 (4): 489-493, 2001

The Coefficient of Covariation

S.M. Seeletse
Department of Statistics, University of South Africa
P.O. Box 392, Pretoria, 0003, South Africa

Abstract: The coefficient of variation (CV) measures the closeness of observed values of a phenomenon, and is limited to data sets of univariate random variables. This article introduces a CV extension to multivariate cases, a measure which we shall call the coefficient of covariation (CCV). The CCV comes in the form of a matrix, the diagonal elements of which are the CVs. The off-diagonal elements of the CCV matrix are CV generalisations derived from covariances and means of the two involved variables. CV measures the closeness of elements of a single data set, and the CCV is intended to measure the closeness of data sets.

Keywords: Coefficient of Variation, Diagonal elements, CCV matrix

Introduction

Let X be a random variable with mean μ and variance σ^2. Denote a random sample of size n from X by $X_1, X_2, \ldots, X_n$, and the respective estimates of μ and σ^2 by $\overline{X}$ and s^2, where

$$\overline{X} = \frac{1}{n}\sum_{k=1}^{n} X_k \qquad (1)$$

$$s^2 = \frac{1}{n-1}\sum_{k=1}^{n}\left(X_k - \overline{X}\right)^2 \qquad (2)$$

Armitage (1971) defines the CV as:

$$c = \frac{s}{\overline{X}} \qquad (3)$$

CV describes situations where changes in the conditions under which change measurement are made alters the standard deviation in the same proportion as it alters the mean (Armitage, 1971). Therefore, CV is a measure of variability that should remain constant. This article generalises the above idea.

Multivariate generalisation: Multivariate generalisation of the random variable X is a random vector $\underline{X} = (X_1\ X_2\ \ldots\ X_p)'$, a p-variate random vector with mean vector $\mu = (\mu_1\ \mu_2\ \ldots\ \mu_p)'$ and a $p \times p$ covariance matrix $\Sigma = \left(\sigma_{ij}\right)$. This means that

$$\mu_i = E\left(X_i\right) \qquad (4)$$

is the population mean of for X_i each i, and

$$\sigma_{ij} = E\left(X_i - \mu_i\right)\left(X_j - \mu_j\right) \qquad (5)$$

is the population covariance between X_i and X_j. We may suppress mention of the term "population". Define

$$\mu_{ij} = \mu_i\mu_j \qquad (6)$$

If $i = j$, then $\sigma_{ii} = \sigma_i^2$, the variance of X_i, and $\mu_{ii} = \mu_i^2$, the square of the mean of X_i. On the other hand, if $i \neq j$, then μ_{ij} is the product of the means of X_i and X_j.

A bivariate random vector has the form (X_i, X_j) and a size n random sample from it is denoted by $(X_{i1}, X_{j1}), (X_{i2}, X_{j2}), \ldots, (X_{in}, X_{jn})$. A size n random sample from $\underline{X}$ is a $p \times n$ matrix $\mathbf{X}$ given by:

$$\mathbf{X} = \begin{bmatrix} X_{11} & X_{12} & . & . & . & X_{1n} \\ X_{21} & X_{21} & . & . & . & X_{21} \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ X_{p1} & X_{p2} & . & . & . & X_{pn} \end{bmatrix} \qquad (7)$$

Generalised forms of (1), (2) and (3), respectively, are given by (8), (9) and (10) in the folowing equations.

$$\overline{X}_i = \frac{1}{n}\sum_{k=1}^{n} X_{ik} \qquad (8)$$

Covariance matrix ($\mathbf{C}$).

$$\begin{array}{c c}
 & [\ldots A_1 \quad \ldots\ldots A_2 \quad \ldots\ldots A_3 \quad \ldots\ldots A_4 \quad \ldots\ldots A_5 \quad \ldots\ldots A_6] \\
\begin{bmatrix} A_1 \\ A_2 \\ A_3 \\ A_4 \\ A_5 \\ A_6 \end{bmatrix} &
\begin{bmatrix}
489.8611 & & & & & \\
50.0833 & 183.25 & & & & \\
-168.278 & -24.208 & 107.944 & & & \\
213.042 & -4.125 & -93.542 & 360 & & \\
-161.208 & 108.5 & 28.083 & -127.375 & 189 & \\
-361.028 & -42.833 & 135.194 & -278.542 & 157.583 & 334.444
\end{bmatrix}
\end{array}$$

The covariance matrix is used to derive the CCV matrix. The diagonal elements of the CCV matrix are the CVs. The CCV matrix is presented next.

Coeficient of covariance matrix.

$$\begin{array}{c c}
 & [\ldots A_1 \quad \ldots\ldots A_2 \quad \ldots\ldots A_3 \quad \ldots\ldots A_4 \quad \ldots\ldots A_5 \quad \ldots\ldots A_6] \\
\begin{bmatrix} A_1 \\ A_2 \\ A_3 \\ A_4 \\ A_5 \\ A_6 \end{bmatrix} &
\begin{bmatrix}
0.490629 & & & & & \\
0.172447 & 0.362598 & & & & \\
-0.29378 & -0.12248 & 0.240377 & & & \\
0.325161 & -0.04974 & -0.22012 & 0.424784 & & \\
-0.34707 & -0.31299 & 0.147991 & -0.31004 & 0.463407 & \\
-0.49837 & -0.19666 & 0.311565 & -0.43992 & 0.406016 & 0.567553
\end{bmatrix}
\end{array}$$

On the question of compatibility, recall that negative correlation implies that one variable increases when the other one decreases, or vice-versa. This means incompatibility. It could be implying clashes of members of a consortium with respect to the attributes and the project. These properties apply also to the CCV. Hence, negative CCV suggests that the two consultancies involved should not partner on the project. (Maybe they could work well together on other projects.)

Findings

- The incompatible pairs for the project in question are given in the order of most conflicting to least conflicting. This list is (A_1, A_6), (A_4, A_6), (A_1, A_5), (A_2, A_5), (A_4, A_5), (A_1, A_3), (A_3, A_4), (A_2, A_6), (A_2, A_3), (A_2, A_4). This means that A_1 conflicts more with A_6 than A_4 is conflicting with A_6, and so on. On the project at hand, elements from the same pair should not be paired together in a consortium for the project. The consortium that is least intolerable is (A_2, A_4) while (A_1, A_6) is the least tolerable.
- Positive CCV entries, listed in the order of most compatible to least compatible, lead to partners (A_5, A_6), (A_1, A_4), (A_3, A_6), (A_1, A_2), (A_3, A_5). It is assumed that if a consultancy has been used in forming a consortium for one project, it may not be used to partner with another firm to compete for the same project. Therefore, by assuming that the first compatibility for a consultancy implies having partnered, then those consultancies appearing for the second time are ignored. Thus, the remaining list of compatibilities becomes (A_5, A_6), (A_1, A_4).

Discussion

The diagonal elements of the CCV matrix are the CVs. The order of CVs from highest stability to lowest leads to the arrangement A_3, A_2, A_4, A_5, A_1, A_6. The consortia of three or more consultancies are difficult to form in general, and particularly when using this approach. For example, in the above example, A_1 is compatible with A_2 and A_4 only, based on their positive CCV entries. But A_2 and A_4 are not compatible because of their negative CCV. Therefore, another justification or criterion should be sought for the possibility of more partners that two in one consortium.

Conclusion and further research: CCV matrix can measure if correlation is negative or positive. It can also measure if pairs of variables are correlated or not, and it can also determine (in)dependence of variables. Further uses of CCV should be researched. It was mentioned that clustering and multicollinearity could be explored using CCV, but details were not given. This hypothesis could also be investigated. Practical situations where CCVs are useful could also be investigated. In the current example, negative CCVs were used to indicate incompatibility, leading to diregarding pairs of the corresponding consultancies.

Appendix

A.1. Mathematical properties of CCV: Intuitively, the statistic c_{ij} is a desirable estimate of the parameter ε_{ij} for each pair i, j. It would be appealing, however, to present mathematical and statistical properties of the CCV matrix and its components. The following theorems and corollaries expose some useful properties of CCV.

Theorem 1

c_{ij} is a desirable estimate of the parameter ε_{ij}

Outline of proof.

Bless & Kathuria (1993) show that for each i, $\overline{X_i}$ is a point estimate of μ_i. Johnson & Wichern (1992) show that S_{ij} is an estimate of σ_{ij}, for each pair i, j. Properties of estimates from Mendenhall et al (1990) and Anderson & Finn (1996) imply that $\frac{s_{ij}}{\overline{\overline{X}}_{ij}}$ is an estimate of $\frac{\sigma_{ij}}{\mu_{ij}}$. (See Appendix A2.)

Further use of Mendenhall et al (1990) leads to the conclusion that $\sqrt{\left|\frac{s_{ij}}{\overline{\overline{X}}_{ij}}\right|} sign(s_{ij})$ is an estimate of $\sqrt{\left|\frac{\sigma_{ij}}{\mu_{ij}}\right|} sign(\sigma_{ij})$, which completes the proof.

Theorem 2

$\varepsilon_{ij} = 0 \Leftrightarrow$ X_i and X_j are uncorrelated.

Proof.

$$\varepsilon_{ij} = \sqrt{\left|\frac{\sigma_{ij}}{\mu_{ij}}\right|} sign(\sigma_{ij})$$

But $\sigma_{ij} = 0 \Leftrightarrow$ X_i and X_j are uncorrelated. Also, $\varepsilon_{ij} = 0 \Leftrightarrow \sigma_{ij} = 0$. Therefore, $\varepsilon_{ij} = 0 \Leftrightarrow$ X_i and X_j are uncorrelated.

The next corollary uses the above theorem on the CCV matrix.

Corollary 3

Ξ is a diagonal matrix $\Leftrightarrow$ the X_i are pairwise uncorrelated.

Proof

The result is accomplished from Theorem 2 by realising that the off-diagonal elements of Ξ are made up of the pairwise elements which become zero when the X_i are pairwise uncorrelated.

The next corollary states a result about independent random variables. Independent variables are automatically uncorrelated, but certain conditions should be satisfied before the converse becomes valid.

Corollary 4

$\varepsilon_{ij} = 0$ if X_i and X_j are independent.

Proof

Independent random variables are uncorrelated (Mendenhall et al, 1990). To complete the proof Theorem 2 can be used.

Corollary 5

Ξ is a diagonal matrix if all the X_i are pairwise independent.

A.2. Properties of estimators

Suppose that a and b are estimators of α and β, respectively. Let $*$ represent any of the operations $+$, $-$, $\times$, or $\div$. Then: $(a+b)^{1/n}$ is an estimator of $(\alpha+\beta)^{1/n}$ for any rational number n, provided that for $* = \div$, $a \neq 0$ and $\alpha \neq 0$.

References

Anderson, T.W. & J. D. Finn. 1996. The new statistical data analysis. Berlin: Springer-Verlag.

Armitage, P., 1971. Statistical methods in medical reserach. Oxford: Blackwell Scientifc Publications.

Bless, C. & R. Kathuria. 1993. Fundamentals of social statistics - an African Perspective. Wetton: Juta & Co.

Johnson, R.A. & D.W. Wichern. 1992. Applied multivariate statistical analysis, 3rd edition. Englewood Cliffs, NJ.: Prentice-Hall.

Mendenhall, W., D.D. Wackerly & R.L. Schaefer., 1990. Mathemetical statistics with aplications, 4th edition. Belmont, California: Duxbury Press.

Meyer, P.L., 1978. Introductory probability and statistical aplications, 2nd edition. London: Addison Wesley.