5-1-2010

# On Exact 100(1-α)% Confidence Interval of Autocorrelation Coefficient in Multivariate Data When the Errors are Autocorrelated

Madhusudan Bhandary

*Columbus State University*, bhandary_madhusudan@columbusstate.edu

# On Exact 100(1-α)% Confidence Interval of Autocorrelation Coefficient in Multivariate Data When the Errors are Autocorrelated

Madhusudan Bhandary
Columbus State University

An exact $100(1-\alpha)\%$ confidence interval for the autocorrelation coefficient $\rho$ is derived based on a single multinormal sample. The confidence interval is the interval between the two roots of a quadratic equation in $\rho$. A real life example is also presented.

Key words: Autocorrelation coefficient, confidence interval, quadratic equation.

## Introduction

The autocorrelation coefficient $\rho$ is frequently used to measure the autocorrelation in a time series model. Weather patterns throughout the year change month by month and there is autocorrelation in the weather pattern from one month to the next. Similarly, the behavior of the stock-market pattern from day to day has an autocorrelation effect.

Statistical inference concerning $\rho$ for a single sample problem has been studied by Durbin and Watson (1950, 1951, 1971); some discussions are also given in Morrison (1983). Cochrane and Orcutt (1949) include a discussion about estimating the regression parameters when the errors are autocorrelated. Bhandary (2005) derived a likelihood ratio test for the equality of two autocorrelation coefficients based on two independent multinormal samples. Bhandary and Doetkott (in review) derived a likelihood ratio test for the equality of more than two autocorrelation coefficients based on more than two independent multinormal samples.

In this article, the problem of developing a method of obtaining an exact $100(1-\alpha)\%$ confidence interval for the autocorrelation coefficient $\rho$ based on a single multinormal

sample is considered. A confidence interval for $\rho$ is found using the distributional property of a statistic. The confidence interval for $\rho$ is the interval between the two roots of a quadratic equation in $\rho$.

## Methodology
### Derivation of the Confidence Interval

The model for the multivariate data with autocorrelated error is as follows:

$$x = \mu + \varepsilon,$$

where

$$x = (x_1 ... x_p)'$$

is a $p \times 1$ vector of observations,

$$\mu = (\mu_1 ... \mu_p)'$$

is a $p \times 1$ vector of unknown means, and

$$\varepsilon = (\varepsilon_1 ... \varepsilon_p)'$$

is a $p \times 1$ vector of random errors.

It is assumed that $\varepsilon \sim N_p(0, \Sigma)$, where

$$\Sigma = \sigma^2 \begin{bmatrix} 1 & \rho & \rho^2 & ... & \rho^{p-1} \\ \rho & 1 & \rho & ... & \rho^{p-2} \\ ... & ... & ... & ... & ... \\ \rho^{p-1} & \rho^{p-2} & ... & ... & 1 \end{bmatrix}, (2.1)$$

Madhusudan Bhandary is an Associate Professor in the Department of Mathematics, Columbus State University, Columbus, GA 31907. Email: bhandary_madhusudan@columbusstate.edu.

and $N_p$ denotes $p$-variate normal distribution. The structure of the covariance matrix in (2.1) means that the errors are autocorrelated. The autocorrelatedness of the error is common in real practice and it can be tested from the data whether the error covariance structure is of (2.1) or not. In expression (2.1), $\sigma^2$ represents the variance of each error component and $\rho$ is called the autocorrelation coefficient.

Let $x_1, x_2, ..., x_n$ be $p \times 1$ vector of $n$ observations independently and identically distributed as $N_p(\mu, \Sigma)$, where $\Sigma$ is given by (2.1). The following transformation can be made: $u_i = T x_i$, $i = 1, 2, ..., n$, where

$$T = \begin{bmatrix} \sqrt{1-\rho^2} & 0 & 0 & .... & 0 \\ -\rho & 1 & 0 & .... & 0 \\ 0 & -\rho & 1 & .... & 0 \\ \multicolumn{5}{c}{\dotfill} \\ 0 & & 0 & 0 & -\rho & 1 \end{bmatrix}. \quad (2.2)$$

Thus,

$$T\Sigma T' = \sigma^2(1-\rho^2)I_p, \quad (2.3)$$

where $I_p$ denotes identity matrix of order $p \times p$. Under the transformation (2.2),

$$u_i \sim N_p\left(\mu^*, \sigma^2(1-\rho^2)I_p\right), \ i = 1, ..., n \quad (2.4)$$

where $\mu^* = T\mu$.

Next, consider splitting the sample data into two parts – one with $n_1$ observations $x_1, x_2, ..., x_{n_1}$ (sample 1) and the other with $n_2$ observations $x_{n_1+1}, x_{n_1+2}, ..., x_{n_1+n_2}$ (sample 2) and $n_1 + n_2 = n$. Using the transformation (2.2), the data vector can be transformed from $x_i$ to $u_i$ as follows:

$$u_i = \begin{pmatrix} u_{i1} \\ u_{i2} \\ . \\ . \\ . \\ u_{ip} \end{pmatrix} \sim N_p\left(\mu^*, \Sigma^*\right); \ i = 1, 2, ..., n \quad (2.5)$$

where $\mu^* = T\mu$, $\Sigma^* = \sigma^2\left(1-\rho^2\right)I_p$ and $T$ is given by (2.2)

Consider the following statistic:

$$\frac{\sum_{i=1}^{n_1}(u_i-\bar{u})'(u_i-\bar{u})}{\sum_{i=n_1+1}^{n} u_i-\bar{u})'(u_i-\bar{u})} \quad (2.6)$$

where $u_i's$ are given by (2.5). Because

$$\frac{(n_2-1)}{(n_1-1)} \frac{\sum_{i=1}^{n_1}(u_i-\bar{u})'(u_i-\bar{u})}{\sum_{i=n_1+1}^{n} (u_i-\bar{u})'(u_i-\bar{u})} \sim F_{n_1 p, n_2 p}$$

(using (2.5)), it can be stated that the exact value of the constant $C^*$ can be obtained from the equation

$$P\left(\frac{\sum_{i=1}^{n_1}(u_i-\bar{u})'(u_i-\bar{u})}{\sum_{i=n_1+1}^{n} u_i-\bar{u})'(u_i-\bar{u})} > C^*\right) = \alpha$$

as

$$C^* = \frac{(n_1-1)}{(n_2-1)} F_{\alpha; n_1 p, n_2 p} \quad (2.7)$$

where, $F_{\alpha; n_1, n_2}$ is the upper $100\alpha\%$ point of the F- distribution with d.f. $n_1, n_2$ respectively and $n_2 = n - n_1$.

The inequality inside the probability in (2.7) can be written as

$$\frac{\sum_{i=1}^{n_1} (x_i - \overline{x})' T'T (x_i - \overline{x})}{\sum_{i=n_1+1}^{n} (x_i - \overline{x})' T'T (x_i - \overline{x})} > C^* \qquad (2.8)$$

where $T$ is given by (2.2).

From (2.2) it may be observed that

$$T'T = \begin{bmatrix} 1 & -\rho & 0 & 0 & \dots & 0 \\ -\rho & 1+\rho^2 & -\rho & 0 & \dots & 0 \\ 0 & -\rho & 1+\rho^2 & -\rho & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & -\rho & 1+\rho^2 & -\rho \\ 0 & 0 & \dots & \dots & -\rho & 1 \end{bmatrix}.$$

$$(2.9)$$

Using (2.9) in (2.8), the following inequality results:

$$\frac{\sum_{i=1}^{n_1} \begin{bmatrix} (x_{i1} - \overline{x}_1)^2 + (1+\rho^2)\sum_{r=2}^{p-1}(x_{ir} - \overline{x}_r)^2 \\ +(x_{ip} - \overline{x}_p)^2 \\ -2\rho\sum_{r=1}^{p-1}(x_{ir} - \overline{x}_r)(x_{\overline{ir+1}} - \overline{x}_{r+1}) \end{bmatrix}}{\sum_{i=n_1+1}^{n} \begin{bmatrix} (x_{i1} - \overline{x}_1)^2 + (1+\rho^2)\sum_{r=2}^{p-1}(x_{ir} - \overline{x}_r)^2 \\ +(x_{ip} - \overline{x}_p)^2 \\ -2\rho\sum_{r=1}^{p-1}(x_{ir} - \overline{x}_r)(x_{\overline{ir+1}} - \overline{x}_{r+1}) \end{bmatrix}} > C^*$$

$$i.e., \frac{\sum_{i=1}^{n_1} \begin{bmatrix} \sum_{r=1}^{p}(x_{ir} - \overline{x}_r)^2 + \rho^2\sum_{r=2}^{p-1}(x_{ir} - \overline{x}_r)^2 \\ -2\rho\sum_{r=1}^{p-1}(x_{ir} - \overline{x}_r)(x_{\overline{ir+1}} - \overline{x}_{r+1}) \end{bmatrix}}{\sum_{i=n_1+1}^{n} \begin{bmatrix} \sum_{r=1}^{p}(x_{ir} - \overline{x}_r)^2 + \rho^2\sum_{r=2}^{p-1}(x_{ir} - \overline{x}_r)^2 \\ -2\rho\sum_{r=1}^{p-1}(x_{ir} - \overline{x}_r)(x_{\overline{ir+1}} - \overline{x}_{r+1}) \end{bmatrix}} > C^*$$

$$i.e. \quad a\rho^2 + b\rho + c > 0 \qquad (2.10)$$

where

$$a = \sum_{i=1}^{n_1}\sum_{r=2}^{p-1}(x_{ir} - \overline{x}_r)^2 - C^*\sum_{i=n_1+1}^{n}\sum_{r=2}^{p-1}(x_{ir} - \overline{x}_r)^2$$

$$b = 2C^*\sum_{i=n_1+1}^{n}\sum_{r=1}^{p-1}(x_{ir} - \overline{x}_r)(x_{\overline{ir+1}} - \overline{x}_{r+1})$$

$$-2\sum_{i=1}^{n_1}\sum_{r=1}^{p-1}(x_{ir} - \overline{x}_r)(x_{\overline{ir+1}} - \overline{x}_{r+1})$$

$$(2.11)$$

and

$$c = \sum_{i=1}^{n_1}\sum_{r=1}^{p}(x_{ir} - \overline{x}_r)^2 - C^*\sum_{i=n_1+1}^{n}\sum_{r=1}^{p}(x_{ir} - \overline{x}_r)^2$$

and $C^*$ is given by (2.7). Note that the data is split and called sample 1 and sample 2 in such a way that a > 0 i.e., if a < 0 then sample 2 is called as sample 1 and sample 1 as sample 2, and in that case a > 0, where a is given by (2.11). (*)

Let the roots of $a\rho^2 + b\rho + c = 0$ be

$$\hat{\rho}_1 = \frac{-b - \sqrt{b^2 - 4ac}}{2a} \qquad (2.12)$$

and

$$\hat{\rho}_2 = \frac{-b + \sqrt{b^2 - 4ac}}{2a}$$

where a, b and c are given by (2.11). Note that in (2.12), if $b^2 - 4ac$ happens to be negative, then sample 2 is called as sample 1 and sample 1 as sample 2 and, in that case, $b^2 - 4ac$ becomes positive.

Because a > 0,

$$\alpha = P(\frac{\sum_{i=1}^{n_1}(u_i - \overline{u})'(u_i - \overline{u})}{\sum_{i=n_1+1}^{n}(u_i - \overline{u})'(u_i - \overline{u})} > C^*)$$

$$= P(a\rho^2 + b\rho + c > 0)$$

$$= P(\rho < \hat{\rho}_1, \rho > \hat{\rho}_2) \text{ (using (*))}.$$

Therefore, $P(\hat{\rho}_1 < \rho < \hat{\rho}_2) = 1 - \alpha$, where $\hat{\rho}_1$ and $\hat{\rho}_2$ are given by (2.12). Thus, $100(1-\alpha)\%$ confidence interval for $\rho$ is $(\hat{\rho}_1, \hat{\rho}_2)$ where $\hat{\rho}_1$ and $\hat{\rho}_2$ are given by (2.12).

A Real Life Example

A real data set from Anderson (1976) and Hand, Daly, Lunn, McConway, and Ostrowski (1994) containing a sample of the monthly average air temperature (°F) from January to April at Nottingham Castle for 10 years - hence the data is autocorrelated over months - is used to provide an example. (See Table 1).

Table1: Average Monthly Temperatures from January to April for Nottingham

| Obsvn. | Year | Month | | | |
|--------|------|-------|------|------|------|
|        |      | Jan   | Feb  | Mar  | Apr  |
| 1      | 1921 | 44.2  | 39.8 | 45.1 | 47.0 |
| 2      | 1937 | 40.8  | 41.0 | 38.4 | 47.4 |
| 3      | 1922 | 37.5  | 38.7 | 39.5 | 42.1 |
| 4      | 1934 | 39.4  | 38.2 | 40.4 | 46.0 |
| 5      | 1927 | 40.6  | 38.5 | 45.3 | 47.1 |
| 6      | 1920 | 36.2  | 40.8 | 44.4 | 46.7 |
| 7      | 1933 | 39.2  | 39.3 | 44.5 | 48.7 |
| 8      | 1926 | 40.0  | 43.4 | 43.4 | 48.9 |
| 9      | 1935 | 40.0  | 42.6 | 43.5 | 47.1 |
| 10     | 1925 | 40.0  | 40.5 | 40.8 | 45.1 |

Results

Data in Table 1 is a p-variate data set for p = 4 and n = 10; it is split into two parts as $n_1 = 5$ and $n_2 = 5$, and sample 1 is the first 5 observations and sample 2 is the remaining 5 observations.

The formulas for computing the sum of square and sum of products are as follows:

$$\left[ \sum_{i=1}^{n_1} \sum_{r=1}^{p} \left( x_{ir}^{(1)} - \bar{x}_r^{(1)} \right)^2 = 92.316 \right]$$

$$\left[ \sum_{i=1}^{n_1} \sum_{r=1}^{p-1} \left( x_{ir}^{(1)} - \bar{x}_r^{(1)} \right)\left( x_{i(r+1)}^{(1)} - \bar{x}_{r+1}^{(1)} \right) = 12.28 \right]$$

$$\left[ \sum_{i=1}^{n_1} \sum_{r=2}^{p-1} \left( x_{ir}^{(1)} - \bar{x}_r^{(1)} \right)^2 = 47.264 \right]$$

$$\left[ \sum_{i=1}^{n_2} \sum_{r=1}^{p} \left( x_{ir}^{(2)} - \bar{x}_r^{(2)} \right)^2 = 41.936 \right]$$

$$\left[ \sum_{i=1}^{n_2} \sum_{r=1}^{p-1} \left( x_{ir}^{(2)} - \bar{x}_r^{(2)} \right)\left( x_{i(r+1)}^{(2)} - \bar{x}_{r+1}^{(2)} \right) = 11.858 \right]$$

$$\left[ \sum_{i=1}^{n_2} \sum_{r=2}^{p-1} \left( x_{ir}^{(2)} - \bar{x}_r^{(2)} \right)^2 = 19.936 \right]$$

Using formula (2.11) results in a = 4.9997, b = 25.7179 and c = 3.4117 and $\alpha = 0.05$ is used, hence $F_{0.05;20,20} = 2.12$. Using (2.12), results in a 95% confidence interval for $\rho$ as (-5.0076, 0.1363) which is approximated as (-1.0, 0.1363) becuase $|\rho| < 1$.

References

Durbin, J., & Watson, G. S. (1950). Testing for serial correlation in least squares regression: I. *Biometrika*, *37*, 409-428.

Durbin, J., & Watson, G. S. (1951). Testing for serial correlation in least squares regression: II. *Biometrika*, *38*, 159-178.

Durbin, J., & Watson, G. S. (1971). Testing for serial correlation in least squares regression: III. *Biometrika*, *58*, 1-19.

Morrison, D. F. (1983). Applied linear statistical methods. *Prentice-Hall,Inc, Englewood Cliffs, NJ.*

Cochrane, D., & Orcutt, G. H. (1949). Application of Least-Squares Regressions to Relationships Containing Auto-correlated error terms. *Journal of the American Statistical Association*, *44*, 32-61.

Bhandary, M. (2005). Test for the equality of autocorrelation coefficients for two populations in multivariate data when the errors are autocorrelated. *Statistics and Probability Letters*, *73*, 333-342.

Bhandary, M., & Doetkott, C. Test for the equality of autocorrelation coefficients for several populations in multivariate data when the errors are autocorrelated. Submitted to *American Journal of Mathematical and Management Science.*

Anderson, O. D. (1976). *Time series analysis and forecasting: The Box-Jenkins approach*. Boston, MA: Butterworth.

Hand, D. J., Daly, F., Lunn, A. D., McConway, K. J., & Ostrowski, E. (1994). *A handbook of small data sets*. London: Chapman & Hall.