PSYCHOMETRIKA—VOL. 71, NO. 3, 529–540
SEPTEMBER 2006
DOI: 10.1007/S11336-04-1221-6

EXACT INTERVAL ESTIMATION, POWER CALCULATION, AND SAMPLE SIZE DETERMINATION IN NORMAL CORRELATION ANALYSIS

GWOWEN SHIEH

NATIONAL CHIAO TUNG UNIVERSITY, TAIWAN

This paper considers the problem of analysis of correlation coefficients from a multivariate normal population. A unified theorem is derived for the regression model with normally distributed explanatory variables and the general results are employed to provide useful expressions for the distributions of simple, multiple, and partial-multiple correlation coefficients. The inversion principle and monotonicity property of the proposed formulations are used to describe alternative approaches to the exact interval estimation, power calculation, and sample size determination for correlation coefficients.

Key words: multiple correlation, partial-multiple correlation, simple correlation.

1. Introduction

Correlation analysis is widely used in many areas of science, and the literature is very extensive. Classical inferences on correlation coefficients are conducted mainly under the assumption that all variables have a joint multivariate normal distribution. Although the underlying normality assumption provides a convenient and useful setup, the resulting probability density functions of the (sample) simple and multiple correlation coefficients r, and R, are notoriously complicated in forms. The complexity incurs continuous investigations to give various expressions, approximations, and computing algorithms for the distributions of both sample correlation coefficients. See Johnson, Kotz, and Balakrishnan (1995, Chap. 32) and Stuart and Ord (1994, Chap. 16) for comprehensive discussions and further details.

The commonly used approximation to the distribution of simple correlation coefficient is Fisher's (1921) z transformation. Several other approximations and asymptotic expansions are described in Johnson et al. (1995, Chap. 32, Secs. 5.2 and 5.3). It appears that the widely used Fisher's z transformation is adequate for moderate sample sizes and the accuracy generally increases with large sample sizes, whereas the other more accurate approximations require more involved computation and/or iterative evaluation. As in the case of simple correlation coefficients, considerable attention has been devoted to the construction of useful approximations for the distribution of the multiple correlation coefficient (see Johnson et al., 1995, Chap. 32, Sec. 11). For the purpose of interval estimation, power calculation, and sample size determination for the squared multiple correlation coefficient, exact results are presented in Algina and Olejnik (2003), Gatsonis and Sampson (1989), Mendoza and Stafford (2001), and Steiger and Fouladi (1992). Although Algina and Olejnik (2003) did not describe their computer algorithms in detail, the exact computations of Gatsonis and Sampson (1989), Mendoza and Stafford (2001), and Steiger and Fouladi (1992) are based on the infinite series expansion of Lee (1972).

The author thanks the referees for their constructive comments and helpful suggestions and especially the associate editor for drawing attention to several critical results which led to substantial improvements of the exposition. The work for this paper was initiated while the author was visiting the Department of Statistics, Stanford University. This research was partially supported by National Science Council Grant NSC-94-2118-M-009-004.

Request for reprints should be sent to Gwowen Shieh, Department of Management Science, National Chiao Tung University, Hsinchu, Taiwan 30050, ROC. E-mail: gwshieh@mail.nctu.edu.tw

In view of the need for evaluating the probabilities of the correlation coefficients and the ultimate aim of presenting exact procedures for correlation analysis, the purpose of this paper is to provide alternative solutions by exploiting the simplification of theoretical property and the accessibility of computing techniques. To this end, a unified theorem is derived for the regression model with multinormal explanatory variables. Although the proposed formulations are based on the intermediate results of multinormal regression and correlation analysis in Anderson (2003), Muirhead (1982), and Sampson (1974), the presentations not only simplify the pedagogical development, but also yield new algorithms for the exact inferences of correlation coefficients. Specifically, the inferential procedures of interval estimation and power calculation in the hypothesis testing situation for simple, multiple, and partial-multiple correlations are described. Furthermore, the planning of sample sizes with estimation and power approaches are also discussed.

In the next section, the major theorem and corollary for the multivariate normal regression model are given. Section 3 applies the proposed formulation to the analysis of the simple correlation coefficient. The presentation is extended to multiple and partial-multiple correlation coefficients in Section 4. Finally, Section 5 contains some concluding remarks.

2. The Multivariate Normal Regression Model

Consider the standard multiple linear regression model with dependent variable Y and all the levels of p independent variables $X(1), \ldots, X(p)$ fixed a priori,

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{1}$$

where $\mathbf{Y} = (Y_1, \ldots, Y_N)^T$, Y_i is the value of the dependent variable Y; $\mathbf{X} = (\mathbf{1}_N, \mathbf{X}_D)$ with $\mathbf{1}_N$ is the $N \times 1$ vector of all 1's, $\mathbf{X}_D = (\mathbf{X}_1, \ldots, \mathbf{X}_N)^{\mathrm{T}}$ is often called the design matrix, $\mathbf{X}_i = (x_{i1}, \ldots, x_{ip})^{\mathrm{T}}$, $x_{i1}, \ldots, x_{ip}$ are the known constants of the p independent variables for $i = 1, \ldots, N$; $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^{\mathrm{T}}$ with $\beta_0, \beta_1, \ldots, \beta_p$ are unknown parameters; and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_N)^{\mathrm{T}}$ with ε_i are independent and indentically distributed as $N(0, \sigma^2)$ random variables. It is well known that under the assumption given above, the likelihood ratio test for the general linear hypothesis $\mathrm{H}_0 : \mathbf{L}\boldsymbol{\beta} = \boldsymbol{\theta}$ versus $\mathrm{H}_1 : \mathbf{L}\boldsymbol{\beta} \neq \boldsymbol{\theta}$ is based on

$$F = \frac{\mathrm{SSH}/l}{\mathrm{SSE}/(N - p - 1)},$$

where $\mathbf{L}$ is an $l \times (p + 1)$ coefficient matrix of rank $l \leq p + 1$, $\boldsymbol{\theta}$ is an $l \times 1$ vector of constants, $\mathrm{SSH} = (\mathbf{L}\hat{\boldsymbol{\beta}} - \boldsymbol{\theta})^{\mathrm{T}}[\mathbf{L}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{L}^{\mathrm{T}}]^{-1}(\mathbf{L}\hat{\boldsymbol{\beta}} - \boldsymbol{\theta})$, $\mathrm{SSE} = (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})^{\mathrm{T}}(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})$, and $b\hat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y}$ is the least squares and maximum likelihood estimator of $\boldsymbol{\beta}$. Under the alternative hypothesis, F is distributed as $F(l, N - p - 1, \Xi)$, the noncentral F-distribution with l and $N - p - 1$ degrees of freedom and noncentrality parameter

$$\Xi = (\mathbf{L}\boldsymbol{\beta} - \boldsymbol{\theta})^{\mathrm{T}}[\mathbf{L}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{L}^{\mathrm{T}}]^{-1}(\mathbf{L}\boldsymbol{\beta} - \boldsymbol{\theta})/\sigma^2.$$

If the null hypothesis is true, then $\Xi = 0$ and F is distributed as $F(l, N - p - 1)$, a central or regular F-distribution with l and $N - p - 1$ degrees of freedom. The test is carried out by rejecting H_0 if $F > F_{l,N-p-1,\alpha}$, where $F_{l,N-p-1,\alpha}$ is the upper α percentage point of the central F-distribution $F(l, N - p - 1)$.

Frequently, the inferences are concerned mainly with the regression coefficients $\boldsymbol{\beta}_1 = (\beta_1, \ldots, \beta_p)^{\mathrm{T}}$ and the corresponding coefficient matrix is written in the form of $\mathbf{L} = \mathbf{L}_1$, where $\mathbf{L}_1 = (\mathbf{0}_c, \mathbf{C})$, $\mathbf{0}_c$ is the $c \times 1$ null vector of all 0's, and $\mathbf{C}$ is a $c \times p$ coefficient matrix of rank $c \leq p$. It follows from the overall estimator $\hat{\boldsymbol{\beta}}$ given above that the prescribed estimator for $\boldsymbol{\beta}_1$ can be expressed as $\hat{\boldsymbol{\beta}}_1 = (\mathbf{X}_C^{\mathrm{T}}\mathbf{X}_C)^{-1}\mathbf{X}_C^{\mathrm{T}})\,\mathbf{Y}$, where $\mathbf{X}_C = (\mathbf{I}_N - \mathbf{J}/N)\mathbf{X}_D$ is the centered form of

$\mathbf{X}_D$, $\mathbf{I}_N$ is the identity matrix of dimension N, and $\mathbf{J}$ is the $N \times N$ square matrix of 1's. With this formulation, it is easily seen that

$$\mathbf{C}\hat{\boldsymbol{\beta}}_1 \sim N_p(\mathbf{C}\beta_1, \boldsymbol{\sigma}^2\mathbf{C}\mathbf{S}_X^{-1}\mathbf{C}^{\mathrm{T}}),$$

where $\mathbf{S}_X = \mathbf{X}_C^{\mathrm{T}}\,\mathbf{X}_C$. Note that $\hat{\sigma}^2 = \mathrm{SSE}/(N - p - 1)$ is the usual unbiased estimator of σ^2 and SSE/σ^2 is distributed as $\chi^2(N - p - 1)$, a chi-square distribution with $N - p - 1$ degrees of freedom and is independent of $\hat{\boldsymbol{\beta}}$. It therefore follows that the general linear hypothesis reduces to $\mathrm{H}_0 : \mathbf{C}\boldsymbol{\beta}_1 = \boldsymbol{\theta}$ versus $\mathrm{H}_1 : \mathbf{C}\boldsymbol{\beta}_1 \neq \boldsymbol{\theta}$ and the F test is conducted by rejecting H_0 if $F^* > F_{c,N-p-1,\alpha}$, where

$$F^* = \frac{\mathrm{SSH}^*/c}{\mathrm{SSE}/(N - p - 1)}, \tag{2}$$

$\mathrm{SSH}^* = (\mathbf{C}\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\theta})^{\mathrm{T}}(\mathbf{C}\mathbf{S}_X^{-1}\mathbf{C}^{\mathrm{T}})^{-1}(\mathbf{C}\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\theta})$. Consequently, F^* is distributed as $F(c, N - p - 1, \Delta)$, where the noncentrality parameter $\Delta = (\mathbf{C}\boldsymbol{\beta}_1 - \boldsymbol{\theta})^{\mathrm{T}}(\mathbf{C}\mathbf{S}_X^{-1}\mathbf{C}^{\mathrm{T}})^{-1}(\mathbf{C}\beta_1 - \theta)/\sigma^2$. Hence, given all model specifications and sample size N, the statistical power achieved for testing hypothesis $\mathrm{H}_0 : \mathbf{C}\boldsymbol{\beta}_1 = \boldsymbol{\theta}$ with specified significance level α against the alternative $\mathrm{H}_1 : \mathbf{C}\boldsymbol{\beta}_1 \neq \boldsymbol{\theta}$ is the probability

$$P\{F(c, N - p - 1, \Delta) > F_{c,N-p-1,\alpha}\}. \tag{3}$$

In the special instance of testing one single coefficient parameter, say $\mathrm{H}_0 : \beta_1 = 0$, it is more flexible to conduct the test with a t statistic since it can be used for one-sided alternatives involving $\mathrm{H}_0 : \beta_1 \leq 0$ or $\mathrm{H}_0 : \beta_1 \geq 0$, while the F statistic cannot. Specifically, the t statistic is

$$t^* = \frac{\hat{\beta}_1}{(\hat{\sigma}^2 s^{11})^{1/2}}, \tag{4}$$

where s^{11} is the (1, 1)th entry of $\mathbf{S}_X^{-1}$ and t^* has a noncentral t distribution $t(N - p - 1, \delta)$ with $N - p - 1$ degrees of freedom and noncentrality parameter $\delta = \beta_1/(\sigma^2 s^{11})^{1/2}$. The corresponding power function is of the form

$$P\{t(N - p - 1, \delta) > t_{N-p-1,\alpha}\} \tag{5}$$

for the one-sided test $\mathrm{H}_0 : \beta_1 \leq 0$ with significance level α, where $t_{N-p-1,\alpha}$ is the upper α percent quantile of the central t-distribution $t(N - p - 1)$, see Rencher (2000, Chaps. 7–8) for further details.

Traditionally, the multiple regression model defined above is referred to as a fixed (conditional) model. The results would be specific to the particular values of the explanatory variables that are observed or preset by the researcher. To extend the concept and applicability of the aforementioned results to the correlation models, the vector of explanatory variables $\{\mathbf{X}_i, i = 1, \ldots, N\}$ in (1) is now assumed to follow a joint multivariate normal distribution with a mean vector $\boldsymbol{\mu}_X$ and a positive definite covariance matrix $\boldsymbol{\Sigma}_X$. It follows immediately from the matrix normal distribution of $\mathbf{X}_D$ that $\mathbf{S}_X$ has a Wishart distribution $W_p(N - 1, \boldsymbol{\Sigma}_X)$. As shown in Sampson (1974, Lemmas 3 and 4), $(\mathbf{C}\mathbf{S}_X^{-1}\mathbf{C}^{\mathrm{T}})^{-1} \sim W_c(N - p + c - 1, (\mathbf{C}\Sigma_X^{-1}\mathbf{C}^{\mathrm{T}})^{-1})$ and, subsequently, $\Delta \sim \Lambda \cdot \chi^2(N - p + c - 1)$ where $\Lambda = (\mathbf{C}\boldsymbol{\beta}_{\ 1} - \boldsymbol{\theta})^{\mathrm{T}}(\mathbf{C}\Sigma_X^{-1}\mathbf{C}^{\ \mathrm{T}})^{-1}(\mathbf{C}\boldsymbol{\beta}_1 - \boldsymbol{\theta})/\boldsymbol{\sigma}^2$. Therefore, the distribution of F^* in the multivariate normal regression model is completely specified in the following theorem.

Theorem 1. *Consider the multiple regression model (1) and $\mathbf{X}_i$ are independent and identically distributed as $N_p(\boldsymbol{\mu}_X, \boldsymbol{\Sigma}_X)$, $i = 1, \ldots, N$. The F^* statistic defined in (2) has the following two-stage distribution*

$$F^*|\Delta \sim F(c, N - p - 1, \Delta) \quad \textit{and} \quad \Delta \sim \Lambda \cdot \chi^2(N - p + c - 1). \tag{6}$$

Note that the formulation (6) also follows from the intermediate results for deriving the density function of R^2 in Anderson (2003, Theorem 4.4.5) and Muirhead (1982, Theorem 5.2.4). However, the expression in Theorem 1 provides a conceptually more transparent representation than those in Theorem 9 and Corollary 2 of Sampson (1974) where the distribution of F^* is expressed as a mixture of central F distributions with random degrees of freedom for the numerator. It is clear under the null hypothesis $\mathrm{H}_0 : \mathbf{C}\boldsymbol{\beta}_1 = \boldsymbol{\theta}$ that $\Lambda = 0$ and Δ degenerates at 0. Hence, the null distribution of F^* remains as $F(c, N - p - 1)$ under both fixed and random settings. However, the power function is more complex than (3) in form due to the extra variability of Δ,

$$P\{F^* > F_{c,N-p-1,\alpha}\} = \int_0^\infty P\{F(c, N - p - 1, \Lambda \cdot K) > F_{c,N-p-1,\alpha}\} \cdot f(K)\, dK, \quad (7)$$

where $f(K)$ is the probability density function of K and $K \sim \chi^2(N - p + c - 1)$. Following similar arguments, it can be shown that the noncentrality δ of the distribution for the t^* statistic defined in (4) has a scaled chi-square distribution $\delta \sim \lambda \cdot \{\chi^2(N - p)\}^{1/2}$, where $\lambda = \beta_1/(\sigma^2\, \sigma^{11})^{1/2}$ and σ^{11} is the (1, 1)th entry of $\boldsymbol{\Sigma}_X^{-1}$. Note that $\sigma^{11}/s^{11} \sim \chi^2(N - p)$. These results are summarized as

Corollary 1. *Consider the multiple regression model (1) and $\boldsymbol{X}_i$ are independent and identically distributed as $N_p(\boldsymbol{\mu}_X, \boldsymbol{\Sigma}_X)$, $i = 1, \ldots, N$. The t^* statistic defined in (4) has the following two-stage distribution*

$$t^*|\delta \sim t(N - p - 1, \delta) \quad \textit{and} \quad \delta \sim \lambda \cdot \{\chi^2(N - p)\}^{1/2}. \quad (8)$$

Thus, the t^* statistic for $\mathrm{H}_0 : \beta_1 \leq 0$ has null distribution $t(N - p - 1)$ and a critical value $t_{N-p-1,\alpha}$ as in the fixed model. Its power can be computed from

$$P\{t^* > t_{N-p-1,\alpha}\} = \int_0^\infty P\{t(N - p - 1, \lambda \cdot \kappa^{1/2}) > t_{N-p-1,\alpha}\} \cdot f(\kappa) d\kappa, \quad (9)$$

where $f(\kappa)$ is the probability density function of κ and $\kappa \sim \chi^2(N - p)$. To exemplify the fundamental differences between the fixed and random model formulations, a direct comparison of the previously defined power functions (5) and (9) shows that the former can be viewed as a realization of the latter based on the observed values of $\mathbf{S}_X$. Consequently, the result would be specific to the particular values of the explanatory variables that are observed in $\mathbf{S}_X$. In another replication of the same study, different settings for the explanatory variables will be obtained. Hence, the conditional power function is not applicable and, more importantly, the fixed modeling approach is not appropriate. The preceding results will be applied later to implement varieties of interval estimation and power calculation in the context of correlation models.

3. Simple Correlation Coefficient

The relation between the multivariate normal regression model and correlation analysis is well known (see Anderson, 2003; Muirhead, 1982; Rencher, 2000). Assume that r is the Pearson product-moment correlation coefficient of (Y_i, X_i), $i = 1, \ldots, N$, where (Y_i, X_i) has a joint bivariate normal distribution $N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_Y \\ \mu_X \end{bmatrix} \quad \text{and} \quad \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_Y^2 & \sigma_{YX} \\ \sigma_{YX} & \sigma_X^2 \end{bmatrix}.$$

The corresponding population correlation coefficient is defined as $\rho = \sigma_{YX}/\sigma_Y\sigma_X$. It follows from standard results that conditional multivariate normal correlation models are equivalent to the usual normal error regression models with the following definitions of notation:

$$\boldsymbol{\beta}_0 = \mu_Y - \rho\mu_X(\sigma_Y/\sigma_X), \quad \beta_1 = \rho(\sigma_Y/\sigma_X), \quad \text{and} \quad \sigma^2 = \sigma_Y^2(1-\rho^2).$$

In the special case of $p = 1$, it is familiar that the reduced t^* statistic can be expressed directly in term of r,

$$t_1 = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}}.$$

Additionally, the test of $\rho \leq 0$ amounts to the test of $\beta_1 \leq 0$ since $\beta_1 = \rho(\sigma_Y/\sigma_X)$. More importantly, it follows from (8) in Corollary 1 that the distribution of t_1 can be represented as

$$t_1|\delta_1 \sim t(N-2, \delta_1) \quad \text{and} \quad \delta_1 \sim \lambda_1 \cdot \{\chi^2(N-1)\}^{1/2},$$

where $\lambda_1 = \rho/(1-\rho^2)^{1/2}$. To demonstrate the discrepancy between the proposed exact formulation and approximate method, and the advantage of the suggested simplifying algorithm, numerical comparisons are conducted to evaluate the widely used Fisher's (1921) z approximation to the distribution function of the simple correlation r. The exact values are computed with the proposed two-stage formulation using programs written with SAS/IML (2003). The results are presented in Table 1 for sample size $N = 10$ and $N = 50$. As expected, the inverse tanh transformation of Fisher (1921) is not sufficiently close to the true distribution of r. However, the performance improves for tail areas and larger sample sizes.

Accordingly, the test of $H_0 : \rho \leq 0$ can be conducted by rejecting H_0 if $t_1 > t_{N-2,\alpha}$. The associated power function is a direct adaptation of (9),

$$P\{t_1 > t_{N-2,\alpha}\} = \int_0^\infty P\{t(N-2, \lambda_1 \cdot \kappa^{1/2}) > t_{N-2,\alpha}\} \cdot f(\kappa)\, d\kappa,$$

where $f(\kappa)$ is the probability density function of κ and $\kappa \sim \chi^2(N-1)$. The numerical computation of exact power requires the evaluation of a noncentral t *cumulative density function* and the one-dimensional integration with respect to a chi-square probability density function. Since all related functions are readily embedded in modern statistical packages such as the SAS system, no substantial computing efforts are required. For the purpose of sample size determination, the minimum sample sizes N required for testing the hypothesis $H_0 : \rho \leq 0$ with a specified parameter value of ρ, significance level, and nominal power, can be found through a simple iterative search. Note that unique and proper solution of the sample size is assured by the monotonicity properties described in Ghosh (1973). The procedures require only obvious modifications for both lower-tailed and two-sided tests.

Interval estimators of ρ can be constructed by the "statistical method" of Mood, Graybill, and Boes (1974, Sec. 4.2) or the "pivoting the *cumulative density function*" method in Casella and Berger (2002, Sec. 9.2.3). For the upper-tailed test just mentioned, the corresponding lower $100(1-\alpha)\%$ confidence interval of ρ is of the form $[-1, \rho_U)$ in which ρ_U (≤ 1) satisfies

$$\int_0^\infty P\{t(N-2, \lambda_{1U} \cdot \kappa^{1/2}) > t_{1O}\} \cdot f(\kappa)\, d\kappa = 1-\alpha,$$

where $\lambda_{1\hat{U}} = \rho U/(1-\rho_U^2)^{1/2}$, $t_{1O} = r_O(N-2)^{1/2}/(1-r_O^2)^{1/2}$, and r_O is the observed value of the simple correlation coefficient. The computations can be easily performed by a standard interval-halving program to meet the desired degree of accuracy. In connection with the interval procedure, it is also critical to ensure adequate estimation accuracy with appropriate sample size. For given values of population correlation coefficient ρ, coverage probability $1-\alpha$, and the

TABLE 1.
The error $= 10^6 \times$ (approximate value $-$ exact value) of Fisher's z approximation to the distribution function of r.

	Cumulative probability												
ρ	0.01	0.05	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	0.95	0.99
$N = 10$													
0.1	794	−2753	−6783	−11285	−12033	−10266	−7037	−3301	−6	1833	1083	−530	−1684
0.2	325	−4462	−9727	−16118	−18140	−17101	−14080	−10047	−5953	−2801	−1686	−2115	−2110
0.3	−162	−6223	−12744	−21032	−24316	−23979	−21136	−16777	−11859	−7382	−4407	−3666	−2524
0.4	−671	−8043	−15843	−26039	−30569	−30907	−28212	−23496	−17730	−11916	−7086	−5187	−2928
0.5	−1206	−9931	−19033	−31147	−36909	−37895	−35315	−30212	−23574	−16408	−9727	−6681	−3323
0.6	−1772	−11898	−22329	−36371	−43347	−44951	−42453	−36931	−29395	−20863	−12333	−8151	−3710
0.7	−2375	−13958	−25747	−41725	−49894	−52085	−49633	−43659	−35200	−25287	−14908	−9599	−4090
0.8	−3027	−16129	−29305	−47226	−56566	−59308	−56865	−50404	−40994	−29683	−17455	−11027	−4462
0.9	−3742	−18436	−33030	−52898	−63379	−66633	−64158	−57173	−46783	−34057	−19977	−12438	−4829
$N = 50$													
0.1	59	−838	−1805	−2973	−3390	−3299	−2879	−2273	−1615	−1041	−697	−627	−442
0.2	−137	−1584	−3073	−4999	−5907	−6094	−5758	−5052	−4103	−3030	−1931	−1348	−629
0.3	−336	−2339	−4354	−7038	−8434	−8894	−8639	−7827	−6583	−5009	−3154	−2060	−813
0.4	−539	−3104	−5647	−9091	−10972	−11701	−11520	−10598	−9055	−6977	−4368	−2766	−994
0.5	−746	−3880	−6955	−11158	−13522	−14515	−14404	−13366	−11520	−8934	−5572	−3464	−1173
0.6	−958	−4666	−8277	−13240	−16084	−17336	−17290	−16131	−13978	−10882	−6767	−4155	−1349
0.7	−1174	−5465	−9613	−15338	−18658	−20165	−20178	−18894	−16428	−12821	−7953	−4840	−1524
0.8	−1394	−6275	−10966	−17452	−21246	−23003	−23070	−21655	−18873	−14751	−9131	−5519	−1696
0.9	−1620	−7098	−12335	−19584	−23848	−25850	−25965	−24414	−21312	−16673	−10300	−6192	−1866

TABLE 2.
The minimum sample sizes required for the prescribed interval $[-1, \rho + b)$ of simple correlation coefficient with coverage probability at least 0.95.

	b			
ρ	0.05	0.10	0.15	0.20
0.00	1084	272	122	69
0.05	1074	269	120	68
0.10	1054	262	117	66
0.15	1023	254	112	63
0.20	982	243	107	60
0.25	932	229	100	56
0.30	874	214	93	52
0.35	808	197	85	47
0.40	736	178	77	42
0.45	658	158	68	37
0.50	578	138	58	32
0.55	495	117	49	26
0.60	411	96	40	21
0.65	330	76	31	16
0.70	252	57	23	12
0.75	180	40	15	8
0.80	117	25	9	NA
0.85	65	13	NA	NA
0.90	26	NA	NA	NA
0.95	NA	NA	NA	NA

bound b (>0), the smallest sample size N required for the sample correlation coefficient to fall into the interval $[-1, \rho + b)$ with probability $1 - \alpha$, is determined by

$$\int_0^\infty P\{t(N-2, \lambda_1 \cdot \kappa^{1/2}) < t_{1U}\} \cdot f(\kappa)\, d\kappa \geq 1 - \alpha,$$

where $\lambda_1 = \rho/(1-\rho^2)^{1/2}$, $t_{1U} = r_U(N-2)^{1/2}/(1-r_U^2)^{1/2}$, and $r_U = \rho + b < 1$. For the purpose of illustration, the minimum sample sizes needed to control the prescribed interval $[-1, \rho + b)$ with coverage probability at least 0.95 are presented in Table 2 for values of ρ ranging from 0 to 0.95 with an increment of 0.05 and $b = 0.05, 0.10, 0.15$, and 0.20. Similarly, the cases of the upper and two-sided $100(1-\alpha)\%$ interval estimation and related sample size calculation can be conducted.

4. Multiple and Partial-Multiple Correlation Coefficients

This section describes the methods for multiple and partial-multiple correlation analysis in the light of the general result given in Theorem 1 for multivariate normal regression models.

4.1. Multiple Correlation Coefficient

Without loss of generality, let $(Y_i, \mathbf{X}_i^{\mathrm{T}})^{\mathrm{T}}$, $i = 1, \ldots, N$, represent the variables in a multivariate correlation model and have a joint $(p+1)$-dimensional multivariate normal distribution

$N_{p+1}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\mathbf{X}_i = (X_{i1}, \ldots, X_{ip})^{\mathrm{T}}$,

$$\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_Y \\ \boldsymbol{\mu}_X \end{bmatrix} \quad \text{and} \quad \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_Y^2 & \boldsymbol{\Sigma}_{YX} \\ \boldsymbol{\Sigma}_{YX}^T & \boldsymbol{\Sigma}_X \end{bmatrix}.$$

One major use of multivariate correlation models is to make inferences on the association between variables Y_i and $\mathbf{X}_i$. A useful measure is the population squared multiple correlation coefficient defined as $\overline{R}^2 = \boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_X^{-1}\boldsymbol{\Sigma}_{YX}^{\mathrm{T}}/\sigma_Y^2$ and the population multiple correlation coefficient $\overline{R}$ is the positive square root of $\overline{R}^2$. The usual sample squared multiple correlation coefficient is denoted by $R^2 = \mathbf{S}_{YX}\mathbf{S}_X^{-1}\mathbf{S}_{YX}^T/s_Y^2$, where $\mathbf{S}_{YX} = \mathbf{Y}^T(\mathbf{I}_N - \mathbf{J}/N)\mathbf{X}_D$ and $s_Y^2 = \mathbf{Y}^T(\mathbf{I}_N - \mathbf{J}/N)\mathbf{Y}$. As in the previous case of simple correlation analysis, the following definitions of notation connect the correlation model of multinormal variables with the multivariate normal regression model: $\beta_0 = \mu_Y - \boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_X^{-1}\boldsymbol{\mu}_X$, $\boldsymbol{\beta}_1 = \boldsymbol{\Sigma}_X^{-1}\boldsymbol{\Sigma}_{YX}^T$, and $\sigma^2 = \sigma_Y^2 - \boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_X^{-1}\boldsymbol{\Sigma}_{YX}^T$. Furthermore, assume the coefficient matrix $\mathbf{C} = \mathbf{I}_p$ and $\boldsymbol{\theta} = \mathbf{0}_p$ in the linear hypothesis of $\mathrm{H}_0 : \mathbf{C}\boldsymbol{\beta}_1 = \boldsymbol{\theta}$, then several simplifications and implications follow from Theorem 1. In particular, Λ turns into $\Lambda_1 = \boldsymbol{\beta}_1^{\mathrm{T}}\boldsymbol{\Sigma}_X\boldsymbol{\beta}_1/\sigma^2 = \overline{R}^2/(1 - \overline{R}^2)$, the population squared multiple correlation coefficient defined above becomes a one-to-one function of the noncentrality Λ_1. This leads to the well-known result that the overall test of regression coefficients $\mathrm{H}_0 : \boldsymbol{\beta}_1 = \mathbf{0}_p$ is equivalent to the test $\mathrm{H}_0 : \overline{R}^2 = 0$. Hence, the inference of $\overline{R}^2$ can be accomplished with the simplified F^* statistic:

$$F_1 = \frac{R^2/p}{(1 - R^2)/(N - p - 1)}$$

and the test $\mathrm{H}_0 : \overline{R}^2 = 0$ is rejected if $F_1 > F_{p,\ N-\ p-1,\alpha}$. It is evident from (6) and (7) that

$$F_1|\Delta_1 \sim F(p, N - p - 1, \Delta_1) \quad \text{and} \quad \Delta_1 \sim \Lambda_1 \cdot \chi^2(N - 1),$$

and the power function of F_1 can be written as

$$P\{F_1 > F_{p,N-p-1,\alpha}\} = \int_0^\infty P\{F(p, N - p - 1, \Lambda_1 \cdot K_1) > F_{p,N-p-1,\alpha}\} \cdot f(K_1)dK_1,$$

where $\Lambda_1 = \bar{R}^2/(1 - \bar{R}^2)$, $f(K_1)$ is the probability density function of K_1 and $K_1 \sim \chi^2(N - 1)$.

For comparative purpose, the suggested simplifying formulation is employed to investigate the accuracy of Lee's (1971, Sec. 5.1) F approximation to the distribution function of R^2 for different values of p and N. Table 3 contains the errors corresponding to Lee's F transformation for $p = 3$ with $N = 10$ and 50. The numerical results suggest that Lee's F transformation for the distribution of R^2 is considerably more accurate than the aforementioned Fisher's z approximation to the distribution of r. To some extent, the performance still varies with the sample size N and the number of parameters p. When $p = 3$ and $N = 10$, there are some cases in Table 3 that give comparatively large errors than other situations. This phenomenon shall continue to exist in other approximations with relatively small p and small N.

The power and sample size calculations can be performed in a similar fashion as in the instance of simple correlation coefficients by the direct substitution of the noncentral t distribution with the noncentral F distribution. It is important to note that the family of noncentral F distributions possesses the same monotonicity properties as those of the family of noncentral t distributions (see Ghosh, 1973).

By pivoting the *cumulative density function*, a $100(1 - \alpha)\%$ one-sided confidence interval of $\overline{R}^2$ in the form of $(0, \overline{R}_U^2)$ can be computed by solving the following equation for $\overline{R}_U^2$:

$$\int_0^\infty P\{F(p, N - p - 1, \Lambda_{1U} \cdot K_1) > F_{1O}\} \cdot f(K_1)\, dK_1 = 1 - \alpha,$$

TABLE 3.
The error $= 10^6 \times$ (approximate value $-$ exact value) of Lee's F approximation to the distribution function of R^2 when $p = 3$.

	Cumulative probability												
$\overline{R}^2$	0.01	0.05	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	0.95	0.99
$N = 10$													
0.1	15	28	28	16	5	−3	−7	−8	−6	−3	−1	0	0
0.2	111	190	167	74	2	−38	−52	−48	−33	−15	−2	1	1
0.3	344	500	387	120	−49	−125	−139	−114	−72	−30	−1	3	1
0.4	701	859	585	103	−153	−248	−243	−187	−111	−42	1	6	2
0.5	1084	1135	689	31	−277	−369	−338	−248	−141	−49	4	9	3
0.6	1352	1256	692	−57	−379	−455	−400	−285	−158	−52	6	12	3
0.7	1406	1215	624	−120	−424	−483	−415	−290	−168	−50	8	12	3
0.8	1214	1020	505	−133	−389	−434	−369	−257	−139	−44	8	11	3
0.9	767	650	323	−88	−255	−286	−244	−171	−93	−30	5	8	2
$N = 50$													
0.1	39	39	18	−14	−28	−30	−24	−13	−3	6	7	4	−0
0.2	77	49	0	−57	−74	−66	−44	−18	6	21	20	10	−1
0.3	92	50	−17	−92	−109	−92	−57	−18	17	37	31	15	−2
0.4	102	50	−31	−119	−136	−111	−66	−16	27	50	39	18	−3
0.5	109	49	−42	−138	−154	−124	−71	−14	34	59	46	20	−4
0.6	110	47	−47	−147	−162	−129	−73	−12	39	64	49	21	−5
0.7	104	44	−47	−143	−157	−125	−70	−10	39	63	48	21	−5
0.8	89	38	−41	−124	−137	−109	−61	−9	34	55	42	18	−4
0.9	59	26	−26	−83	−92	−74	−42	−7	22	37	29	13	−3

TABLE 4.
The minimum sample sizes required for the prescribed interval $[0, \overline{R}^2 + b)$ of squared multiple correlation coefficient with coverage probability at least 0.95 and $p = 5$.

	b			
$\overline{R}^2$	0.05	0.10	0.15	0.20
0.00	221	110	73	55
0.05	414	154	90	63
0.10	551	184	101	68
0.15	649	204	108	70
0.20	714	215	111	71
0.25	749	219	111	70
0.30	757	217	108	67
0.35	744	210	103	63
0.40	711	197	96	58
0.45	662	182	87	53
0.50	600	163	78	46
0.55	529	142	67	40
0.60	451	120	57	33
0.65	371	98	46	27
0.70	291	76	35	20
0.75	214	56	25	14
0.80	145	37	17	NA
0.85	85	21	NA	NA
0.90	39	NA	NA	NA
0.95	NA	NA	NA	NA

where $\Lambda_{1U} = \bar{R}^2_U/(1 - \bar{R}^2_U)$, $F_{1O} = \{(N - p - 1)/p\}\{R^2_O/(1 - R^2_O)\}$, and R^2_O is the observed value of the squared multiple correlation coefficient. However, proper positive values of $\overline{R}^2_U$ are found only if $F_{1O} > F_{p,N-p-1,1-\alpha}$. Additionally, it is of interest to consider the planning of sample sizes for interval estimation with the prescribed length and desired accuracy. With the specified quantities of population squared multiple correlation coefficient $\overline{R}^2$, target probability $1 - \alpha$, and the bound b (>0), the minimum sample size N required for the interval $[0, \bar{R}^2 + b)$ with coverage probability at least $1 - \alpha$ can be computed from

$$\int_0^\infty P\{F(p, N - p - 1, \Lambda_1 \cdot K_1) < F_{1U}\} \cdot f(K_1)\, dK_1 \geq 1 - \alpha,$$

where $\Lambda_1 = \overline{R}^2/(1 - \overline{R}^2)$, $F_{1U} = \{(N - p - 1)/p\}\{\overline{R}^2_U/(1 - \overline{R}^2_U)\}$, and $\overline{R}^2_U = \overline{R}^2 + b < 1$. For demonstration, the minimum sample sizes needed to guarantee the prescribed interval $[0, \overline{R}^2 + b)$ with coverage probability at least 0.95 and $p = 5$ are presented in Table 4 for $\overline{R}^2$ ranges from 0 to 0.95 with an increment of 0.05 and $b = 0.05, 0.10, 0.15$, and 0.20. Furthermore, the extensions for the upper and two-sided $100(1 - \alpha)\%$ interval estimation and related sample size determination are straightforward.

4.2. Partial-Multiple Correlation Coefficient

Another problem of particular interest is the analysis of population squared partial-multiple correlation $\overline{R}^2_{2.1}$ between Y and $X(p - q + 1), \ldots, X(p)$ after controlling $X(1), \ldots, X(p - q)$ where $p > q > 0$. To fix the idea, the p variables $\mathbf{X}_i$ are divided into two sets $\mathbf{X}_{1i}$ and $\mathbf{X}_{2i}$ of size

$p - q$ and q, respectively. Use the following notation for partitioning the corresponding arrangement of the matrices

$$\boldsymbol{\Sigma}_{YX} = \begin{bmatrix} \boldsymbol{\Sigma}_{Y1} & \boldsymbol{\Sigma}_{Y2} \end{bmatrix} \quad \text{and} \quad \boldsymbol{\Sigma}_{X} = \begin{bmatrix} \boldsymbol{\Sigma}_{X1} & \boldsymbol{\Sigma}_{X12} \\ \boldsymbol{\Sigma}_{X12}^{\mathrm{T}} & \boldsymbol{\Sigma}_{X2} \end{bmatrix}.$$

Furthermore, define

$$\begin{bmatrix} \sigma_{Y.1}^2 & \boldsymbol{\Sigma}_{Y2.1} \\ \boldsymbol{\Sigma}_{Y2.1}^{\mathrm{T}} & \boldsymbol{\Sigma}_{X2.1} \end{bmatrix} = \begin{bmatrix} \sigma_{Y}^2 & \boldsymbol{\Sigma}_{Y2} \\ \boldsymbol{\Sigma}_{Y2}^{\mathrm{T}} & \boldsymbol{\Sigma}_{X2} \end{bmatrix} - \begin{bmatrix} \boldsymbol{\Sigma}_{Y1} \\ \boldsymbol{\Sigma}_{X12}^{\mathrm{T}} \end{bmatrix} \boldsymbol{\Sigma}_{X1}^{-1} [\boldsymbol{\Sigma}_{Y1}^{T} \boldsymbol{\Sigma}_{X12}].$$

Then, it follows that $\overline{R}_{2.1}^2 = \boldsymbol{\Sigma}_{Y2.1} \boldsymbol{\Sigma}_{X2.1}^{-1} \boldsymbol{\Sigma}_{Y2.1}^{T} / \sigma_{Y.1}^2$. According to the definition of $\boldsymbol{\beta}_1 = \boldsymbol{\Sigma}_{X}^{-1} \boldsymbol{\Sigma}_{YX}^{T}$ given before, its last q components can be written as $\boldsymbol{\beta}_2 = \boldsymbol{\Sigma}_{X2.1}^{-1} \boldsymbol{\Sigma}_{Y2.1}^{\mathrm{T}}$. Then $\boldsymbol{\Sigma}_{Y2.1} \boldsymbol{\Sigma}_{X2.1}^{-1} \boldsymbol{\Sigma}_{Y2.1}^{\mathrm{T}} = \boldsymbol{\beta}_2^{\mathrm{T}} \boldsymbol{\Sigma}_{X2.1} \boldsymbol{\beta}_2 = \sigma^2 (\overline{R}^2 - \overline{R}_1^2)/(1 - \overline{R}^2)$ and $\sigma_{Y.1}^2 = \sigma^2 (1 - \overline{R}_1^2)/(1 - \overline{R}^2)$, where $\bar{R}_1^2$ is the population squared multiple correlation coefficient between variables Y and $X(1), \ldots, X(p - q)$. Hence, the hypothesis testing of $\mathrm{H}_0 : \overline{R}_{2.1}^2 = 0$ is equivalent to the one of $\mathrm{H}_0 : \boldsymbol{\beta}_2 = \mathbf{0}_q$, where the last test can be expressed in the form of linear hypothesis $\mathrm{H}_0 : \mathbf{C}\boldsymbol{\beta}_1 = \boldsymbol{\theta}$ with $c = q$, $\mathbf{C} = [\mathbf{0}_{q \times (p - q)}, \mathbf{I}_q]$, and $\boldsymbol{\theta} = \mathbf{0}_q$, where $\mathbf{0}_{q \times (p - q)}$ is a $q \times (p - q)$ matrix of all 0's. As an illustration of the general F^* statistic defined in (4), the resulting partial F statistic is

$$F_2 = \frac{(R^2 - R_1^2)/q}{(1 - R^2)/(N - p - 1)},$$

where R_1^2 is the sample squared multiple correlation coefficient between Y and the first $p - q$ independent variables $X(1), \ldots, X(p - q)$. The distribution of F_2 follows as a direct consequence of Theorem 1 that $F_2 | \Delta_2 \sim F(q, N - p - 1, \Delta_2)$ and $\Delta_2 \sim \Lambda_2 \cdot \chi^2 (N - p + q - 1)$, where $\Lambda_2 = (\overline{R}^2 - \overline{R}_1^2/(1 - \overline{R}^2) = \overline{R}_{2.1}^2/(1 - \overline{R}_{2.1}^2)$. Hence, the test $\mathrm{H}_0 : \overline{R}_{2.1}^2 = 0$ is rejected if $F_2 > F_{q,\ N-p-1,\alpha}$. The power function becomes

$$P\{F_2 > F_{q,N-p-1,\alpha}\} = \int_0^\infty P\{F(q, N - p - 1, \Lambda_2 \cdot K_2) > F_{q,N-p-1,\alpha}\} \cdot f(K_2)\, dK_2,$$

where $f(K_2)$ is the probability density function of K_2 and $K_2 \sim \chi^2 (N - p + q - 1)$.

It is noteworthy that the strong resemblances between the distributions and power functions of F_1 and F_2 for the tests of multiple and partial-multiple correlation coefficients, respectively. Fundamentally, the inferences of the partial-multiple correlation coefficient can be conducted in a similar manner as presented in the previous section for the multiple correlation coefficient. The details are not provided here.

5. Conclusions

This paper presents a simplified treatment of multivariate normal regression models that are tied to correlation models with multinormal variables. A full range of exact methods for correlation analysis is then considered. The proposed results are notable in the conceptual clarity of formulations for the well-known but complicated distributions of simple, multiples and partial-multiple correlations. Consequently, the suggested procedures provide alternative approaches to perform normal correlation analysis in conjunction with basic computation techniques that require only standard numerical methods of one-dimensional integration and an interval-halving algorithm. The integration is theoretically exact provided that the auxiliary functions can be evaluated exactly. The essential part involves the auxiliary functions of noncentral t and F and central χ^2 distributions. The SAS/IML codes for carrying out the computation of the proposed methods are available from the website: www.ms.nctu.edu.tw/faculty/shieh.

References

Algina, J., & Olejnik, S. (2003). Sample size tables for correlation analysis with applications in partial correlation and multiple regression analysis. *Multivariate Behavioral Research*, *38,* 309–323.

Anderson, T.W. (2003). *An introduction to multivariate statistical analysis* (3rd ed.). New York: Wiley.

Casella, G., & Berger, R.L. (2002). *Statistical inference* (2nd ed.). Pacific Grove, CA: Duxbury.

Fisher, R.A. (1921). On the probable error of a coefficient of correlation deduced from a small sample. *Metron*, *1,* 3–32.

Gatsonis, C., & Sampson, A.R. (1989). Multiple correlation: Exact power and sample size calculations. *Psychological Bulletin*, *106,* 516–524.

Ghosh, B.K. (1973). Some monotonicity theorems for χ^2, F and t distributions with applications. *Journal of the Royal Statistical Society*, *Series B*, *35,* 480–492.

Johnson, N.L., Kotz, S., & Balakrishnan, N. (1995). *Continuous univariate distributions* (2nd ed., Vol. 2). New York: Wiley.

Lee, Y.S. (1971). Some results on the sampling distribution of the multiple correlation coefficient. *Journal of the Royal Statistical Society*, *Series B*, *33,* 117–129.

Lee, Y.S. (1972). Tables of upper percentage points of the multiple correlation coefficient. *Biometrika*, *59,* 175–189.

Mendoza, J.L., & Stafford, K.L. (2001). Confidence interval, power calculation, and sample size estimation for the squared multiple correlation coefficient under the fixed and random regression models: A computer program and useful standard tables. *Educational and Psychological Measurement*, *61,* 650–667.

Mood, A.M., Graybill, F.A., & Boes, D.C. (1974). *Introduction to the theory of statistics* (3rd ed.). New York: McGraw-Hill.

Muirhead, R.J. (1982). *Aspects of multivariate statistical theory*. New York: Wiley.

Rencher, A.C. (2000). *Linear models in statistics*. New York: Wiley.

Sampson, A.R. (1974). A tale of two regressions. *Journal of the American Statistical Association*, *69,* 682–689.

SAS Institute (2003). *SAS/IML user's guide*, *Version 8*. Carey, NC: author.

Steiger, J.H., & Fouladi, R.T. (1992). R2: A computer program for interval estimation, power calculations, sample size estimation, and hypothesis testing in multiple regression. *Behavioral Research Methods, Instruments, and Computers*, *24,* 581–582.

Stuart, A., & Ord, J. K. (1994). *Kendall's advanced theory of statistics* (6th ed., Vol. 1). New York: Halsted Press.

Manuscript received 29 JUN 2004
Final version received 24 JUN 2005
Published Online Date: 25 AUG 2006