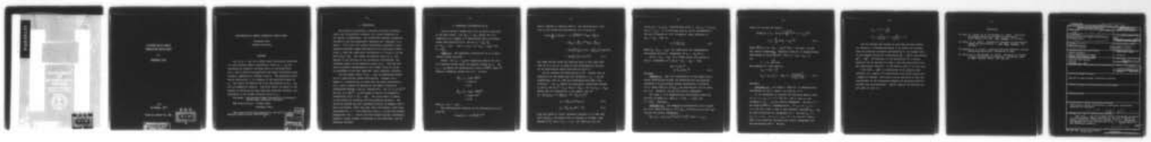
AD-A048 126   CLEMSON UNIV S C DEPT OF MATHEMATICAL SCIENCES   F/G 12/1
DISTRIBUTION OF SAMPLE CORRELATION COEFFICIENTS.(U)
SEP 77 K ALAM   N00014-75-C-0451
UNCLASSIFIED   N88   NL

1 OF 1
ADA048 126



END
DATE
FILMED
1 -78
DDC

DISTRIBUTION OF SAMPLE CORRELATION COEFFICIENTS

BY

KHURSHEED ALAM

N88

SEPTEMBER, 1977

TECHNICAL REPORT No. 266

DISTRIBUTION OF SAMPLE CORRELATION COEFFICIENTS

Khursheed Alam

Clemson University

ABSTRACT

Let $(Y, X_1, \ldots, X_K)$ be a random vector distributed according to a multivariate normal distribution, where $X_1, \ldots, X_K$ are considered as predictor variables and Y is the predictand. Let r_i and R_i denote the population and sample correlation coefficients, respectively, between Y and X_i. The population correlation coefficient r_i is a measure of the predictive power of X_i. The author has derived the joint distribution of $R_1, \ldots, R_K$ and its asymptotic property. The given result is useful in the problem of selecting the most important predictor variable, corresponding to the largest absolute value of r_i.

Key Words: Multivariate Normal Distribution; Correlation Coefficients; Predictor Variables.

AMS Classification: Primary 62E15

Secondary 62E20

*The author's work was supported by the Office of Naval Research under Contract N00014-75-C-0451.

1. INTROUCTION

The problem of selecting a variable or several variables from a set of predictor variables $\{X_i\}$ occurs frequently in the design of experiments. The correlation between a predictor variable X_i and the predictand Y measures the "leverage" of X_i upon Y. If X_i and Y are jointly distributed according to the standard bivariate normal distribution with correlation coefficient r_i then the conditional distribution of Y given X_i is normal $N(r_i X_i, 1-r_i^2)$. The larger the absolute value of r_i, the smaller is the variance of the conditional distribution, and therefore higher is the predictive power of X_i. Thus the predictor variable corresponding to the largest value of r_i^2 may be considered as the most important (best) predictor variable.

Let the random vector $(Y, X_1, \ldots, X_K)$ be distributed according to a multivariate normal distribution. Suppose that a sample of n observations is taken from the given distribution. Let r_i and R_i denote the population and sample correlation coefficients between Y and X_i, respectively. Let $r_i^* = r_i(1-r_i^2)^{-\frac{1}{2}}$ and $R_i^* = R_i(1-R_i^2)^{-\frac{1}{2}}$. In this paper we derive the asymptotic distribution of $\underset{\sim}{R}^* = (R_1^*, \ldots, R_K^*)$. The given result is useful in the problem of selecting the best predictor variable. The selection problem has been considered recently by Ramberg (1976). Rizvi and Solomon (1976) and Alam, Rizvi and Solomon (1976) have considered the ...m of selecting from $p \geq 2$ given multivariate populations, the ...lation with the largest multiple correlation between a single variate, classified as the predictand, and the remaining variates.

2. ASYMPTOTIC DISTRIBUTION OF $\underset{\sim}{R}^*$

First we prove a lemma which will be used in the proof of the main result. Let $(Z_{1t},\ldots,Z_{Kt})$ denote the t-th observation in a sample of n observations from a K-variate normal distribution $N(\underset{\sim}{0},\Omega)$, and let $U_i = \sum_{t=1}^{n} Z_{it}^2$, $\underset{\sim}{U} = (U_1,\ldots,U_K)'$. Let $\Omega = (\omega_{ij})$, $\underset{\sim}{e} = (\omega_{11},\ldots,\omega_{KK})'$ and $\hat{\Omega} = (\omega_{ij}^2)$.

<u>Lemma 2.1</u>. The asymptotic distribution of $\underset{\sim}{U}$ is multivariate normal $N(n\underset{\sim}{e},2n\hat{\Omega})$.

Proof: Let $\theta_1,\ldots,\theta_K$ be K imaginary numbers and let D be a diagonal matrix whose ith diagonal element is θ_i. Let $\lambda_1,\ldots,\lambda_K$ denote the characteristic roots of $\Omega^{\frac{1}{2}}D\Omega^{\frac{1}{2}}$, where $\Omega^{\frac{1}{2}}$ denotes a symmetric square root of Ω. We have

$$\begin{aligned}\sum_{i=1}^{K} \lambda_i &= \text{trace } \Omega^{\frac{1}{2}}D\Omega^{\frac{1}{2}}\\ &= \text{trace } D\Omega\\ &= \underset{\sim}{\theta}'\underset{\sim}{e}\end{aligned}$$

$$\begin{aligned}\sum_{i=1}^{K} \lambda_i^2 &= \text{trace } (\Omega^{\frac{1}{2}}D\Omega^{\frac{1}{2}})^2\\ &= \text{trace } (D\Omega)^2\\ &= \underset{\sim}{\theta}'\hat{\Omega}\underset{\sim}{\theta}\end{aligned}$$

where $\underset{\sim}{\theta} = (\theta_1,\ldots,\theta_K)'$.

The characteristic function of the distribution of $\underset{\sim}{U}$ is given by

$$E\exp(\underset{\sim}{\theta}'\underset{\sim}{U}) = |I-2\Omega^{\frac{1}{2}}D\Omega^{\frac{1}{2}}|^{-n/2}$$

where I denotes an identity matrix. The characteristic function of the normalized distribution of $\underset{\sim}{U}$ is given by

$$E \exp(\frac{1}{\sqrt{2n}} \underset{\sim}{\theta}'(\underset{\sim}{U}-n\underset{\sim}{e})) = |I-\sqrt{\tfrac{2}{n}}\Omega^{\frac{1}{2}}D\Omega^{\frac{1}{2}}|^{-n/2}\exp(-\sqrt{\tfrac{2}{n}}\underset{\sim}{\theta}'\underset{\sim}{e})$$

$$= \Pi_{i=1}^{K}(1-\sqrt{\tfrac{2}{n}}\lambda_i)^{-n/2}\exp(-\sqrt{\tfrac{2}{n}}\underset{\sim}{\theta}'\underset{\sim}{e})$$

$$= \exp\sqrt{\tfrac{2}{n}}(\textstyle\sum_{i=1}^{K}\lambda_i-\underset{\sim}{\theta}'\underset{\sim}{e}) + \tfrac{1}{2}\textstyle\sum_{i=1}^{K}\lambda_i^2)(1+O(n^{-\frac{1}{2}}))$$

$$= \exp(\tfrac{1}{2}\underset{\sim}{\theta}'\hat{\Omega}\underset{\sim}{\theta})(1+O(n^{-\frac{1}{2}})). \qquad (2.1)$$

The lemma follows since the limiting value of the right hand side of (2.1) as $n\to\infty$ is equal to the characteristic function of the multivariate normal distribution $N(\underset{\sim}{0},\hat{\Omega})$.□

Now we consider the distribution of $\underset{\sim}{R}^*$. Without loss of generality we can assume that the variables $Y, X_1,\ldots,X_K$ are standarized, that is, they are distributed with mean 0 and variance 1. Let r_{ij} denote the correlation coefficient between X_i and X_j and let $\Sigma = (r_{ij})$ and $\hat{\Sigma} = (r_{ij}^2)$. Let $(Y_t,X_{1t},\ldots,X_{Kt})$ denote the t-th observation in the sample, and let

$\overline{X}_i = \frac{1}{n}\sum_{t=1}^{n}X_{it}$, $\overline{Y} = \frac{1}{n}\sum_{t=1}^{n}Y_t$, $S^2 = \sum_{t=1}^{n}(Y_t-\overline{Y})^2$

$$V_i = (\textstyle\sum_{t=1}^{n}(Y_t-\overline{Y})X_{it})/S \qquad (2.2)$$

$$W_i = \textstyle\sum_{t=1}^{n}(X_{it}-\overline{X})^2 - V_i^2. \qquad (2.3)$$

From the theory of linear regression analysis it is seen that $W_i\overset{d}{\sim}(1-r_i^2)\chi^2_{n-2}$, chi-square with n-2 degrees of freedom, independent of V_i and $\underset{\sim}{Y} = (Y_1,\ldots,Y_n)'$ and $V_i\overset{d}{\sim}N(r_iS,1-r_i^2)$ and

$\operatorname{cov}(V_i, V_j) = r_{ij} - r_i r_j$, conditionally given $\underset{\sim}{Y}$. Let $\lambda_{ij} = r_{ij} - r_i r_j$ and $\Omega = (\lambda_{ij})$. It is also seen that W_i can be represented as the sum of squares of (n-2) orthogonal linear functions of $X_{i1}, \ldots, X_{in}$. That is

$$W_i \overset{d}{=} \sum_{t=1}^{n-2} Z_{it}^2 \qquad (2.4)$$

Where $\underset{\sim}{Z}_t = (Z_{1t}, \ldots, Z_{kt})'$ are identically and independently distributed as $N(\underset{\sim}{0}, \Omega)$, independent of $V_1, \ldots, V_K$ and S.

Let $\underset{\sim}{T} = (T_1, \ldots, T_K)'$ be a random vector distributed as $N(\underset{\sim}{0}, \Omega)$, independent of S and $\underset{\sim}{W} = (W_1, \ldots, W_K)'$. Then

$$\begin{aligned} R_i^* &= V_i (W_i)^{-\frac{1}{2}} \\ &\overset{d}{\approx} (T_i + r_i S) W_i^{-\frac{1}{2}}. \end{aligned} \qquad (2.5)$$

Therefore

<u>Theorem 2.1</u>. The joint distribution of the sample correlation coefficients between the predictand and the predictor variables of a multivariate normal distribution is given by (2.5), where $\underset{\sim}{T} \overset{d}{\approx} N(0, \Omega)$, $S^2 \overset{d}{\approx} \chi^2_{n-1}$, the distribution of $\underset{\sim}{W}$ is given by (2.4). Moreover, $(S, \underset{\sim}{T}, \underset{\sim}{W})$ are jointly independent.

For large n, $\underset{\sim}{W}$ is asymptotically distributed as $N((n-2)\underset{\sim}{f}, 2(n-2)\hat{\Omega})$ by Lemma 2.1, where $\underset{\sim}{f} = (1-r_1^2, \ldots, 1-r_K^2)'$ and $\hat{\Omega} = (\lambda_{ij}^2)$. Therefore

<u>Corollary 2.1</u>. The asymptotic distribution of $\underset{\sim}{R}^*$ is given by (2.5), where $\underset{\sim}{T} \overset{d}{\approx} N(0, \Omega)$, $S^2 \overset{d}{\approx} \chi^2_{n-1}$, $W \overset{d}{\approx} N((n-2)f, 2(n-2)\hat{\Omega})$ and $(S, \underset{\sim}{T}, \underset{\sim}{W})$ are jointly independent.

Let $\gamma_{ij} = (r_{ij} - r_i r_j)^2 (1-r_i^2)^{-1} (1-r_j^2)^{-1}$ and $\Gamma = (\gamma_{ij})$.

From (2.5) we have for large n

$$\sqrt{n}(R_i^*-r_i^*) = T_i + \sqrt{n}\, r_i^*\left(\left(\frac{(1-r_i^2)S^2}{W_i}\right)^{\frac{1}{2}} - 1\right) + O_p(n^{-\frac{1}{2}})$$

$$= T_i + \frac{r_i^*}{\sqrt{2}}(A-B_i) + O_p(n^{-\frac{1}{2}}) \qquad (2.6)$$

Where $A \overset{d}{\sim} N(0,1)$, $\underset{\sim}{B} = (B_1,\ldots,B_K)' \overset{d}{\approx} N(0,\Gamma)$. Moreover, $(\underset{\sim}{T},A,\underset{\sim}{B})$ are jointly independent. Therefore $\sqrt{n}(\underset{\sim}{R}^*-\underset{\sim}{r}^*)$ is asymptotically distributed as $N(\underset{\sim}{0},\Omega+\frac{r_i^{*2}}{2}(\Gamma+E))$ where $E = (e_{ij})$, $e_{ij} = 1$.

Let

$$C = \Omega + \frac{r_i^{*2}}{2}(\Gamma+E).$$

The elements of C are given by

$$C_{ii} = 1-r_i^2 + r_i^{*2}$$

$$C_{ij} = r_{ij}-r_ir_j + r_i^*r_j^*\left[1 + \frac{(r_{ij}-r_ir_j)^2}{(1-r_i^2)(1-r_j^2)}\right]. \qquad (2.7)$$

Therefore

<u>Corollary 2.2</u>. For large n, $\sqrt{n}(\underset{\sim}{R}^*-\underset{\sim}{r}^*)$ is asymptotically distributed as $N(\underset{\sim}{0},C)$, where C is given by (2.7).

It is interesting to consider the following special cases: (1) $r_i = 0$, $r_{ij} = 0$, $i \neq j$ for all i and j, that is, the variables $Y, X_1,\ldots,X_K$ are jointly independent. We have $C = I$ and $\sqrt{n}(\underset{\sim}{R}^*-\underset{\sim}{r}^*) \overset{d}{\sim} N(0,I)$, asymptotically. (2) $r_i = 0$, $r_{ij} = \rho$, $i \neq j$ for all i and j, that is the predictor variables $X_1,\ldots,X_K$ are equi-correlated and independent of Y. We have $C_{ii} = 1$, $C_{ij} = \rho$, $i \neq j$. (3) $r_i = \rho$, $r_{ij} = 0$, $i \neq j$ for all i and j, that is the predictor variables are jointly independent and equi-correlated with Y. We have

$$c_{ii} = 1-\rho^2 + \frac{\rho^2}{1-\rho^2}$$

$$c_{ij} = \frac{\rho^2}{2(1-\rho^2)}\left(1 + \frac{\rho^4}{(1-\rho^2)^2}\right) - \rho^2.$$

Now we consider the problem of selecting the best predictor variable. A standard procedure is to select the variable from the predictor variables corresponding to the largest value of the squared correlation coefficients $R_1^2,\ldots,R_K^2$ or equivalently $R_1^{*2},\ldots,R_K^{*2}$. By Corollary 2.2 the probability of a correct selection can be derived for large n from the multivariate normal distribution function. In the special case (1) we have that

n max $(R_1^{*2},\ldots,R_K^{*2})$ is distributed as the largest order statistic in a sample of K observations from χ_1^2-chi-square with 1 degree of freedom. This result can be used also to test the significance of the correlation between the selected predictor variable and the predictand. Similar results are obtained for the cases (2) and (3).

References

[1] Alam, K., Rizvi, M. H. and Solomon, H. (1976). Selection of largest multiple correlation coefficients: exact sample size case. Ann. Statist. (4) 614-620.

[2] Gupta, S. S. (1963). Probability integrals of multivariate normal and multivariate t. Ann. Math. Statist. (34) 792-828.

[3] Ramberg, J. S. (1977). Selecting the best Predictor variate. Unpublished Manuscript, University of Iowa.

[4] Rizvi, M. H. and Solomon, H. (1973). Selection of largest multiple correlation coefficients: asymptotic case. J. Amer. Statist. Assoc. (68) 184-188.

UNCLASSIFIED
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered) 14 N88 TR-266

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: N-88
2. GOVT ACCESSION NO.:
3. RECIPIENT'S CATALOG NUMBER:
4. TITLE (and Subtitle): 6 Distribution of Sample Correlation Coefficients.
5. TYPE OF REPORT & PERIOD COVERED: 9 Technical rept.
6. ~~PERFORMING ORG. REPORT NUMBER~~
7. AUTHOR(s): 10 Khursheed/Alam
8. CONTRACT OR GRANT NUMBER(s): 15 N00014-75-C-0451
9. PERFORMING ORGANIZATION NAME AND ADDRESS:
Clemson University
Dept. of Mathematical Sciences
Clemson, South Carolina 29631
10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: NR 042-271
11. CONTROLLING OFFICE NAME AND ADDRESS:
Office of Naval Research
Code 436
Arlington, Va. 22217
12. REPORT DATE: 11 Sep 77
13. NUMBER OF PAGES: 12 11p.
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office):
15. SECURITY CLASS. (of this report): Unclassified
15a. DECLASSIFICATION/DOWNGRADING SCHEDULE:

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Multivariate normal distribution; correlation coefficient; selection of variables.

sub 1 sub K 407 183 sub i

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Let $(Y, X_1, \dots, X_K)$ be a random vector distributed according to a multivariate normal distribution, and let R_i denote the sample correlation coefficient between Y and X_i. ~~The author has derived~~ the joint distribution of $R_1, \dots, R_K$ and ~~has~~ given its asymptotic property.

sub i sub 1 are derived sub K

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE
S/N 0102-014-6601

UNCLASSIFIED
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)