Netherlands School of Economics
ECONOMETRIC INSTITUTE

Report 6902

THE CORRELATION COEFFICIENT IN THE GENERAL LINEAR MODEL

by J. Koerts and A.P.J. Abrahamse

January 10, 1969

Preliminary and Confidential

THE CORRELATION COEFFICIENT IN THE GENERAL LINEAR MODEL

by J. Koerts and A.P.J. Abrahamse

Contents

1. INTRODUCTION

In economic research, one frequently uses the so-called general linear model. It can be written in matrix notation as

$$y = X\beta + u \tag{1.1}$$

where y is a column vector of n values taken by the dependent variable, X a matrix of order $n \times \Lambda$ of nonstochastic values taken by the Λ explanatory variables, β a column vector of Λ unknown parameters, and u a vector of n normally distributed random variables. It is important to distinguish this model from the regression model. In the regression model, the X-matrix consists of particular values of random variables while in the linear model, the X-matrix is composed of nonstochastic or mathematical variables. This difference may seem to be slight but it has some important consequences. In the regression model, for instance, there is generally no functional relationship between the variables while the main hypothesis of the linear model is that a functional relationship exists even though it may be disturbed by random errors. The strong resemblance between these two models coupled

[1] The authors want to thank Mr. A.S. Louter of the Econometric Institute for writing the needed computer programs and his participation in several discussions.

with the fact that the same terminology has been introduced in both models has caused confusion among econometricians. One important example occurs in the regression model where quantity known as the correlation coefficient has been defined. It is considered as a measure of interdependency between random variables. In the general linear model, a similar quantity has been introduced also under the name correlation coefficient. One is, of course, free to define such a quantity, but problems arise if one starts using it as if it were a "true" correlation coefficient. That is, as if the variables involved were random variables when at least some of them are nonstochastic.

The purpose of this paper is to study the problems arising from attempts to draw inferences by using the correlation coefficient in the general linear model. To be more precise: we wish to consider whether or not it is possible in the general linear model to use the correlation coefficient as a test variable for the correct specification of an economic law. With respect to the specification given in (1.1), throughout this paper we shall use the classical assumptions:

(1) The X-matrix consists of nonstochastic elements and has full-column rank.

(2) The vector u is normally distributed with mean vector zero and covariance matrix $\sigma^2 I$.

The order of the discussion runs as follows. In section 2, the correlation coefficient is defined and its probability limit is determined. In section 3, the distribution of the correlation coefficient is (theoretically) derived. In section 4, a number of examples are discussed and the influence of the X-matrix on the distribution of the correlation coefficient is considered. In section 5, some final conclusions are drawn.

2. DEFINITIONS AND PRELIMINARY RESULTS

2.1. The Linear Model with one Explanatory Variable

As we have already observed the correlation coefficient is often used in the search for the "correct" specification of an economic law. Theoretical considerations also play an important role in the solution of this choice problem. From the class of theoretically acceptable specification, that with the highest correlation coefficient is usually chosen. In order to examine this strategy in more detail, we consider the linear model with one explanatory variable and a constant term

$$y_i = \alpha + \beta x_i + u_i \qquad i = 1, \ldots, n \tag{2.1}$$

where α and β are unknown parameters, x the value taken by the explanatory variable,

y the value taken by the dependent variable, and u the disturbance term. In this model, the correlation coefficient is defined by

$$R = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (y_i - \bar{y})^2}} \tag{2.2}$$

where $\bar{x}$ denotes the average of the values taken by the explanatory variable and $\bar{y}$ the sample mean of the dependent variable. The squared correlation coefficient can be written as follows

$$R^2 = 1 - \frac{\sum_{i=1}^{n} v_i^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \tag{2.3}$$

where v_i for $i = 1, \ldots, n$ stands for the i^{th} least-squares estimated disturbance. This can be proved as follows:

$$\sum_{i=1}^{n} v_i^2 = \sum_{i=1}^{n} [y_i - bx_i - a]^2 = \sum_{i=1}^{n} [y_i - bx_i - (\bar{y} - b\bar{x})]^2 \tag{2.4}$$

where a and b are the least-squares estimates of the parameters α and β respectively. If we define

$$X_i = x_i - \bar{x} \qquad Y_i = y_i - \bar{y} \tag{2.5}$$

we can write, using $b = \sum X_i Y_i / \sum X_i^2$,

$$\begin{aligned} \sum_{i=1}^{n} v_i^2 &= \sum_{i=1}^{n} [Y_i - bX_i]^2 = \sum_{i=1}^{n} Y_i^2 - 2b \sum_{i=1}^{n} X_i Y_i + b^2 \sum_{i=1}^{n} y_i^2 \\ &= \sum_{i=1}^{n} Y_i^2 - 2 \frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} X_i^2} \sum_{i=1}^{n} X_i Y_i + \frac{(\sum_{i=1}^{n} X_i Y_i)^2}{(\sum_{i=1}^{n} X_i^2)^2} \sum_{i=1}^{n} X_i^2 \\ &= \sum_{i=1}^{n} Y_i^2 - \frac{(\sum_{i=1}^{n} X_i Y_i)^2}{\sum_{i=1}^{n} X_i^2} = (1 - R^2) \sum_{i=1}^{n} Y_i^2 \end{aligned} \tag{2.6}$$

which proves (2.3).

In order to get more insight into the behavior of R^2, we determine the probability limit of the correlation coefficient. From (2.3) and (2.1) we have

$$R^2 = 1 - \frac{\frac{1}{n}\sum_{i=1}^{n} v_i^2}{\frac{1}{n}\beta^2 \sum_{i=1}^{n} (x_i - \bar{x})^2 + \frac{1}{n}\sum_{i=1}^{n} (u_i - \bar{u})^2} \tag{2.7}$$

because

$$y_i - \bar{y} = \beta(x_i - \bar{x}) + (u_i - \bar{u}) \tag{2.8}$$

and the cross product is equal to zero.

Before we can determine the probability limit of R^2, we must make an assumption about the behavior of the explanatory variable x for large values of n. One usually assumes that $\sum_{i=1}^{n} (x_i - \bar{x})^2/n$ remains bounded for increasing values of n and tends to a limit S^2. This assumption is a sufficient condition for the least-squares estimates of α and β to be consistent.[2] Let us now determine the probability limit of R^2.

Firstly, consider the numerator of the quotient given in (2.7) which will be denoted by $s^2 = \Sigma v_i^2/n$. Notice that

$$\frac{ns^2}{\sigma^2} = \frac{y'My}{\sigma^2} = \frac{u'Mu}{\sigma^2} = z'Mz \tag{2.9}$$

where M is the well-known projection matrix $[I - X(X'X)^{-1}X']$, z a vector of normal variables with zero mean and unit variance, and where use is made of the idempotency of M and the property that $MX = 0$. Thus the quantity $z'Mz$ is chi-square distributed with $n - 2$ degrees of freedom, $n - 2$ being the rank of M. And thus

$$E[s^2] = \frac{\sigma^2}{n} E[z'Mz] = \frac{\sigma^2}{n}(n - 2) \tag{2.10}$$

---

[2] Of course, the well-known assumptions about the vector of disturbances u are also needed. Moreover, it is not necessary for $\sum_i (x_i - \bar{x})^2/n$ to tend to a limit, it only should be bounded.

and

$$\text{(2.11)} \qquad \operatorname{var} s^2 = \frac{\sigma^4}{n^2} \operatorname{var} (z'Mz) = \frac{\sigma^4}{n^2} 2(n - 2)$$

From (2.10) and (2.11), it follows that, for large n, the expectation of s^2 approaches σ^2 and the variance of s^2 approaches zero, which implies that $\sum_{i=1}^{n} v_i^2/n$ converges in the quadratic mean to σ^2.

Secondly, we consider the probability limit of the denominator of (2.7). By assumption, we have

$$\text{(2.12)} \qquad \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 = S^2$$

Moreover,

$$\text{(2.13)} \qquad \frac{1}{n} \sum_{i=1}^{n} (u_i - \bar{u})^2 = \frac{1}{n} \sum_{i=1}^{n} u_i^2 - \bar{u}^2$$

According to the law of large numbers, we have

$$\text{(2.14)} \qquad \operatorname{plim} \frac{1}{n} \sum_i u_i^2 = \sigma^2$$

and

$$\operatorname{plim} \bar{u}^2 = 0$$

Hence we can conclude that

$$\text{(2.15)} \qquad \operatorname{plim} R^2 = 1 - \frac{\sigma^2}{\beta^2 S^2 + \sigma^2} = \frac{\beta^2 S^2}{\beta^2 S^2 + \sigma^2}$$

The following conclusions can now be drawn: (1) the probability limit of R^2 is always smaller than 1 and - depending on the values of S^2 and β - it can deviate substantially from 1; (2) a positive relationship exists between plim R^2 and the absolute value of β; (3) if $\Sigma (x_i - \bar{x})^2/n$ is not bounded but becomes infinite for n approaching infinity, the probability limit of R^2 is equal to 1. A decreasing trend in the observations on the explanatory variable tends to lower plim R^2.

2.2. The Linear Model with more than one Explanatory Variable

We now generalize the results of the previous section. To do this, we must introduce the multiple correlation coefficient. This quantity is defined as the simple correlation coefficient between the dependent variable y and the least-squares prediction of y, which will be denoted by y^*. In matrix notation:

$$y^* = Xb \tag{2.16}$$

where X denotes the matrix of values taken by the explanatory variables and b the vector of least-squares estimates of the vector of unknown parameters β_i.

If we write

$$Y_i^* = y_i^* - \bar{y}^* \tag{2.17}$$

where $\bar{y}^*$ is the average of the values taken by y^*, the multiple correlation coefficient is defined as

$$R = \frac{\sum_{i=1}^{n} Y_i Y_i^*}{\sqrt{\sum_{i=1}^{n} Y_i^2 \sum_{i=1}^{n} y_i^{*2}}} \tag{2.18}$$

In much the same way as for the simple correlation coefficient, it can be shown that the squared multiple correlation coefficient is equal to

$$R^2 = 1 - \frac{\sum_{i=1}^{n} v_i^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \tag{2.19}$$

where v_i for $i = 1, \ldots, n$ denotes the i^{th} least-squares estimated disturbance. To prove this, we first consider the numerator of (2.18).

$$\sum_{i=1}^{n} Y_i Y_i^* = \sum_{i=1}^{n} (Y_i^* + v_i) Y_i^* = \sum_{i=1}^{n} Y_i^{*2} + \sum_{i=1}^{n} Y_i^* v_i = \sum_{i=1}^{n} Y_i^{*2} \tag{2.20}$$

because

$$\begin{aligned} \sum_{i=1}^{n} Y_i^* v_i &= \sum_{i=1}^{n} [b_1 X_{1i} + b_2 X_{2i} + \ldots + b_\Lambda X_{\Lambda i}] v_i \\ &= b_1 \sum_{i=1}^{n} X_{1i} v_i + b_2 \sum_{i=1}^{n} X_{2i} v_i + \ldots + b_\Lambda \sum_{i=1}^{n} X_{\Lambda i} v_i = 0 \end{aligned} \tag{2.21}$$

where the explanatory variables are measured in deviation of their respective means. The last equality sign of (2.21) follows directly from the normal equations and the fact that a constant term is added in the specification. Next, we rewrite the $\sum_{i=1}^{n} v_i^2$ using (2.20)

$$(2.22) \qquad \sum_{i=1}^{n} v_i^2 = \sum_{i=1}^{n} (Y_i - Y_i^*)^2 = \sum_{i=1}^{n} Y_i^2 - 2 \sum_{i=1}^{n} Y_i Y_i^* + \sum_{i=1}^{n} Y_i^{*2}$$

$$= \sum_{i=1}^{n} Y_i^2 - \sum_{i=1}^{n} Y_i Y_i^* = \frac{\sum_{i=1}^{n} Y_i^2 - \sum_{i=1}^{n} Y_i Y_i^*}{\sum_{i=1}^{n} Y_i^2} \sum_{i=1}^{n} Y_i^2$$

and using (2.20) and (2.18) we get

$$(2.23) \qquad \sum_{i=1}^{n} v_i^2 = (1 - R^2) \sum_{i=1}^{n} Y_i^2 = (1 - R^2) \sum_{i=1}^{n} (y_i - \bar{y})^2$$

from which (2.19) follows immediately.

We shall now determine the probability limit of the multiple correlation coefficient. We first write (2.19) in matrix notation

$$(2.24) \qquad R^2 = 1 - \frac{\sum_{i=1}^{n} v_i^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} = 1 - \frac{y'My}{y'[I - \frac{1}{n}\iota\iota']y} = 1 - \frac{u'Mu}{y'Ey}$$

where we have used the fact that the vector of least-squares residuals equals My. In (2.24), I is the identity matrix and ι is a column vector with unity for each of its elements. The matrix E, which is of course equal to $[I - \frac{1}{n}\iota\iota']$, is an operator which measures a variable as deviation of its mean. Using the relation

$$(2.25) \qquad y = Xb + v$$

we can write (2.24) as

$$(2.26) \qquad R^2 = 1 - \frac{u'Mu}{(Xb + v)'E(Xb + v)} = 1 - \frac{u'Mu}{b'(X'EX)b + v'Ev + 2v'EXb}$$

$$= 1 - \frac{u'Mu}{b'(X'EX)b + v'Ev} = 1 - \frac{u'Mu}{b'(X'EX)b + v'v} =$$

$$= 1 - \frac{u'Mu}{b'(X'EX)b + u'Mu}$$

where we use the fact that $v'EXb = 0$, which follows from the normal equations (see (2.21)).

In order to compute the probability limit of R^2, we have to make an assumption about the behavior of the matrix (X'EX). We assume that the matrix (X'EX)/n remains bounded for large n and tends to a non-singular matrix A. This assumption is sufficient for the least-squares estimator b of β to be consistent. It is a straightforward generalization of that made in the case of the simple correlation coefficient. We are now ready to determine the probability limit of R^2. We have already pointed out that the numerator of (2.26) u'Mu/n converges to σ^2 in probability and, with regard to the denominator, we have by assumption

$$\text{plim } b'(X'EX)b/n = \beta'A\beta \tag{2.27}$$

The probability limit of the multiple correlation coefficient is thus equal to

$$\text{plim } R^2 = \text{plim } \left[1 - \frac{\frac{1}{n}u'Mu}{\frac{1}{n}b'(X'EX)b + \frac{1}{n}u'Mu}\right] = 1 - \frac{\sigma^2}{\beta'A\beta + \sigma^2} = \frac{\beta'A\beta}{\beta'A\beta + \sigma^2} \tag{2.28}$$

Notice that (2.28) is a straightforward generalization of (2.15). From (2.28) we can conclude that the plim of R^2 is always smaller than 1.

It is of interest to observe that besides the value of σ^2 the quadratic form $\beta'A\beta$ becomes important. The degree of singularity of A in particular will influence the probability limit of R^2.

3. THE DERIVATION OF THE DISTRIBUTION OF R^2

In this section, we describe how the distribution function of R^2 can be determined by means of numerical integration. In other words: we are looking for the distribution function

$$F(r^2) = P[R^2 \leq r^2] \tag{3.1}$$

where

$$R^2 = 1 - \frac{y'My}{y'Ey} = \frac{y'Ey - y'My}{y'Ey} = \frac{y'(E - M)y}{y'Ey} \tag{3.2}$$

y being an n-element column vector of normally distributed random variables with mean vector $X\beta$ and covariance matrix $\sigma^2 I$. Substitution of (3.2) into (3.1) gives

$$(3.3) \qquad F(r^2) = P\left[\frac{y'(E - M)y}{y'Ey} \leq r^2\right] = P[y'(E - M)y \leq r^2 y'Ey]$$

$$= P[y'\{(1 - r^2)E - M\}y \leq 0]$$

$$= P[y'(kE - M)y \leq 0]$$

where k is equal to $1 - r^2$.

Thus the problem of finding the distribution of a ratio of quadratic forms is reduced simply to that of a mere quadratic form. The matrix $kE - M$ is symmetric, so it can be diagonalized by an appropriate orthogonal transformation as

$$(3.4) \qquad P'(kE - M)P = D$$

where D is a diagonal matrix with characteristic roots on its main diagonal and P is an orthogonal matrix of characteristic vectors. Let us next make the orthogonal transformation $z = P'y/\sigma$. The quadratic form in (3.3) then reduces to

$$(3.5) \qquad y'(kE - M)y = y'(PDP')y = \sigma^2 z'Dz = z'\Lambda z$$

where z is a vector of normally distributed random variables with mean vector $P'X\beta/\sigma$ and covariance matrix I, and $\Lambda = \sigma^2 D$. Hence (3.3) can be written as

$$(3.6) \qquad F(r^2) = P[z'\Lambda z \leq 0] = P\left[\sum_{j=1}^{n} \lambda_j z_j^2 \leq 0\right]$$

where $\Lambda = \{\text{diag. } \lambda_j\}$,[3] and where the z_j^2 are independent noncentral chi-square variables, each with one degree of freedom and a non-centrality parameter δ_j^2 which equals the squared j^{th} element of the vector $P'X\beta/\sigma$.

Making use of a result discovered by Imhof [1], we have

$$(3.7) \qquad F(r^2) = P\left[\sum_{j=1}^{n} \lambda_j z_j^2 \leq 0\right] = \frac{1}{2} - \frac{1}{\pi}\int_0^{\infty} \frac{\sin \varepsilon(u)}{u\gamma(u)}\, du$$

where

$$(3.8) \qquad \varepsilon(u) = \tfrac{1}{2}\sum_{j=1}^{n} [tg^{-1}(\lambda_j u) + \delta_j^2 \lambda_j u(1 + \lambda_j^2 u^2)^{-1}] - \tfrac{1}{2} r^2 u$$

---

[3] The λ_j are equal to the characteristic roots of $kE - M$ up to a multiplicative constant σ^2.

and

$$\gamma(u) = \prod_{j=1}^{n} (1 + \lambda_j^2 u^2)^{\frac{1}{4}} \exp \tfrac{1}{2}\{\sum_{j=1}^{n} \delta_j^2 \lambda_j^2 u^2/(1 + \lambda_j^2 u^2)\} \tag{3.9}$$

The integral in (3.7) can be calculated by numerical integration making use of

$$\lim_{u\to 0} \frac{\sin \varepsilon(u)}{u\gamma(u)} = \tfrac{1}{2} \sum_{j=1}^{n} \lambda_j (1 + \delta_j^2)$$

The function $u\gamma(u)$ increases monotonically towards infinity. Therefore, in numerical work, the integration will be carried out over a finite range $0 \leq u \leq U$. Hence two different types of error exist: (1) the error arising from using an approximate rule to determine the integral, and (2) a truncation error. The latter $t_u = \frac{1}{\pi} \int_U^\infty \frac{\sin \varepsilon(u)}{u\gamma(u)}\, du$ has an upper bound T_u

$$|T_u| \leq \frac{1}{\pi} \int_U^\infty \left|\frac{1}{u\gamma(u)}\right| du$$

so this error can be controlled and thus $F(r^2)$ can be computed to any desired degree of accuracy.

From this discussion it is clear that both the X-matrix and the value of σ^2 influence the distribution of R^2. We study these influences in the next section.

4. SOME EXAMPLES AND THE INFLUENCE OF THE X-MATRIX

4.1. The Textile Example

Our first example deals with the demand for textiles in the Netherlands during a time period of 15 years between the two world wars, thus $n = 15$. There are two explanatory variables: the logarithm of real income per head and the logarithm of the deflated price index of the commodity. A constant term is added and therefore $\Lambda = 3$. For the vector of parameters β, we use estimates of the constant term, and the income and price elasticities based on the available data.

Using the method described in section 3, we compute the distribution function of R^2 for different values of σ^2, given the X-matrix and the vector β. The results can be seen in Table 1. It is interesting to see that, for values of σ^2 equal to 1 or 0.1, nearly all the probability mass is concentrated in the interval $[0, 0.5]$ For $\sigma = 0.01$, the interval is somewhat larger, though almost all the probability mass still lies below 0.8. The influence of σ^2 is quite clear; the smaller the values of σ^2, the more the probability density function of R^2 is shifted towards the

right. This tendency can also be seen from the pictures on page 13. It is easy to understand this influence because σ appears in the denominator of the non-centrality parameters δ_j^2. The smaller σ^2 the larger δ_j^2 for $j = 1, \ldots, n$.

TABLE 1

r^2	$\sigma^2 = 1$ $F(r^2)$	$\sigma^2 = 0.1$ $F(r^2)$	$\sigma^2 = 0.01$ $F(r^2)$
0.0000	0.0000	0.0000	0.0000
0.0416	0.2173	0.1565	0.0057
0.0833	0.3943	0.2982	0.0166
0.1250	0.5371	0.4244	0.0342
0.1666	0.6510	0.5353	0.0599
0.2083	0.7408	0.6310	0.0952
0.2500	0.8107	0.7124	0.1408
0.2916	0.8643	0.7823	0.1970
0.3333	0.9047	0.8358	0.2636
0.3750	0.9347	0.8803	0.3395
0.4166	0.9564	0.9152	0.4226
0.4583	0.9718	0.9418	0.5103
0.5000	0.9824	0.9614	0.5992
0.5416	0.9894	0.9756	0.6854
0.5833	0.9937	0.9853	0.7652
0.6250	0.9968	0.9916	0.8352
0.6666	0.9984	0.9956	0.8926
0.7083	0.9993	0.9979	0.9364
0.7500	0.9997	0.9991	0.9668
0.7916	0.9999	0.9997	0.9854
0.8333	0.9999	0.9999	0.9949
0.8750	0.9999	0.9999	0.9988
0.9166	0.9999	0.9999	0.9999
0.9583	0.9999	0.9999	0.9999
1.0000	1.0000	1.0000	1.0000

The consequence of this result is important for it implies that, given the specific explanatory variables used (the X-matrix) and given the assumptions made about the vector of random disturbances u, it is very unlikely that a correlation coefficient higher than 0.8 will be found. In other words, if the vector y is really generated by the given X-matrix and a vector of disturbances u, which is normally distributed with mean vector zero and covariance matrix $\sigma^2 I$, then a correlation coefficient higher than 0.8 is very unlikely. It is perhaps interesting to mention that the correlation coefficient found in practice was 0.987. This result is very unlikely in the light of the foregoing discussion and perhaps is an indication that one has looked for explanatory variables which, for the <u>given</u> set of y values, yield a high correlation coefficient. In any case, the assumptions seem to be in conflict with oneanother unless one wants to accept a very small value of σ^2.

The question now arises of why the probability distribution of R^2 is concentrated so much at the lower values of R^2. We discuss this interesting question in the next section.

4.2. An Artificial Example

We now consider an artificial example in an attempt to trace the influence of the X-matrix on the probability distribution of R^2. We therefore construct another X-matrix of the same order as the one used in the textile example.[4] The ranges of the values taken by the explanatory variables are chosen in such a way that they are approximately equal to the ranges of the variables used in the textile example. Moreover, we use the same vector of parameters β. The corresponding distributions of R^2, for different values of σ^2, are computed and given in Table 2.

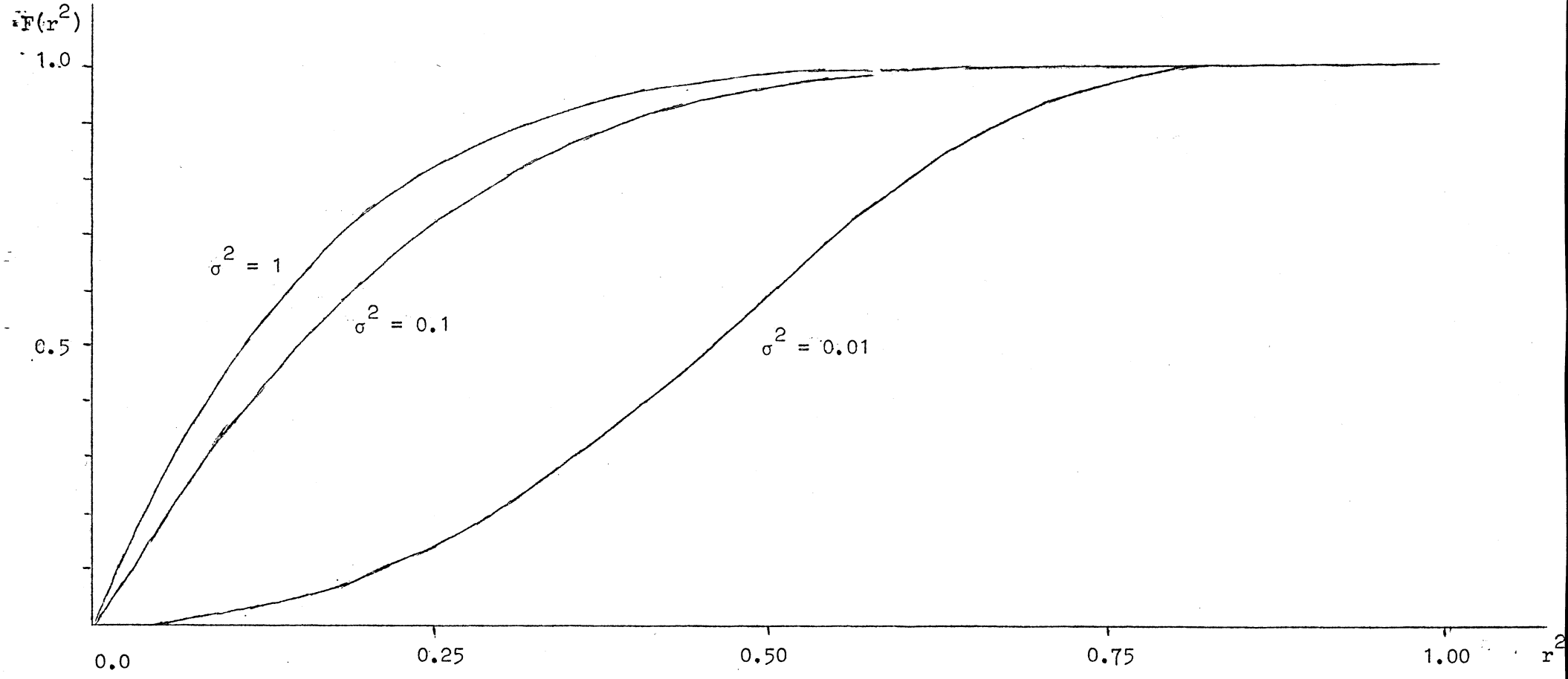
The differences between Tables 1 and 2 are very striking. The distribution of R^2 is now completely shifted towards the right, that is, the probability mass is concentrated at higher values of R^2. This example shows how sensitive the distribution of R^2 is to changes in the X-matrix. Moreover, the influence of a change in σ^2 seems to have an even more pronounced effect.

TABLE 2

	$\sigma^2 = 1$	$\sigma^2 = 0.1$	$\sigma^2 = 0.01$
r^2	$F(r^2)$	$F(r^2)$	$F(r^2)$
0.0000	0.0000	0.0000	0.0000
0.0416	0.0007	0.0000	0.0000
0.0833	0.0026	0.0000	0.0000
0.1250	0.0066	0.0000	0.0000
0.1666	0.0137	0.0000	0.0000
0.2083	0.0254	0.0000	0.0000
0.2500	0.0434	0.0000	0.0000
0.2916	0.0696	0.0000	0.0000
0.3333	0.1059	0.0000	0.0000
0.3750	0.1540	0.0000	0.0000
0.4166	0.2150	0.0000	0.0000
0.4583	0.2892	0.0000	0.0000
0.5000	0.3755	0.0000	0.0000
0.5416	0.4713	0.0000	0.0000
0.5833	0.5726	0.0000	0.0000
0.6250	0.6736	0.0000	0.0000
0.6666	0.7682	0.0000	0.0000
0.7083	0.8499	0.0000	0.0000
0.7500	0.9140	0.0003	0.0000
0.7916	0.9583	0.0027	0.0000
0.8333	0.9841	0.0199	0.0000
0.8750	0.9959	0.1155	0.0000
0.9166	0.9995	0.4471	0.0000
0.9583	0.9998	0.9170	0.0000
1.0000	1.0000	1.0000	1.0000

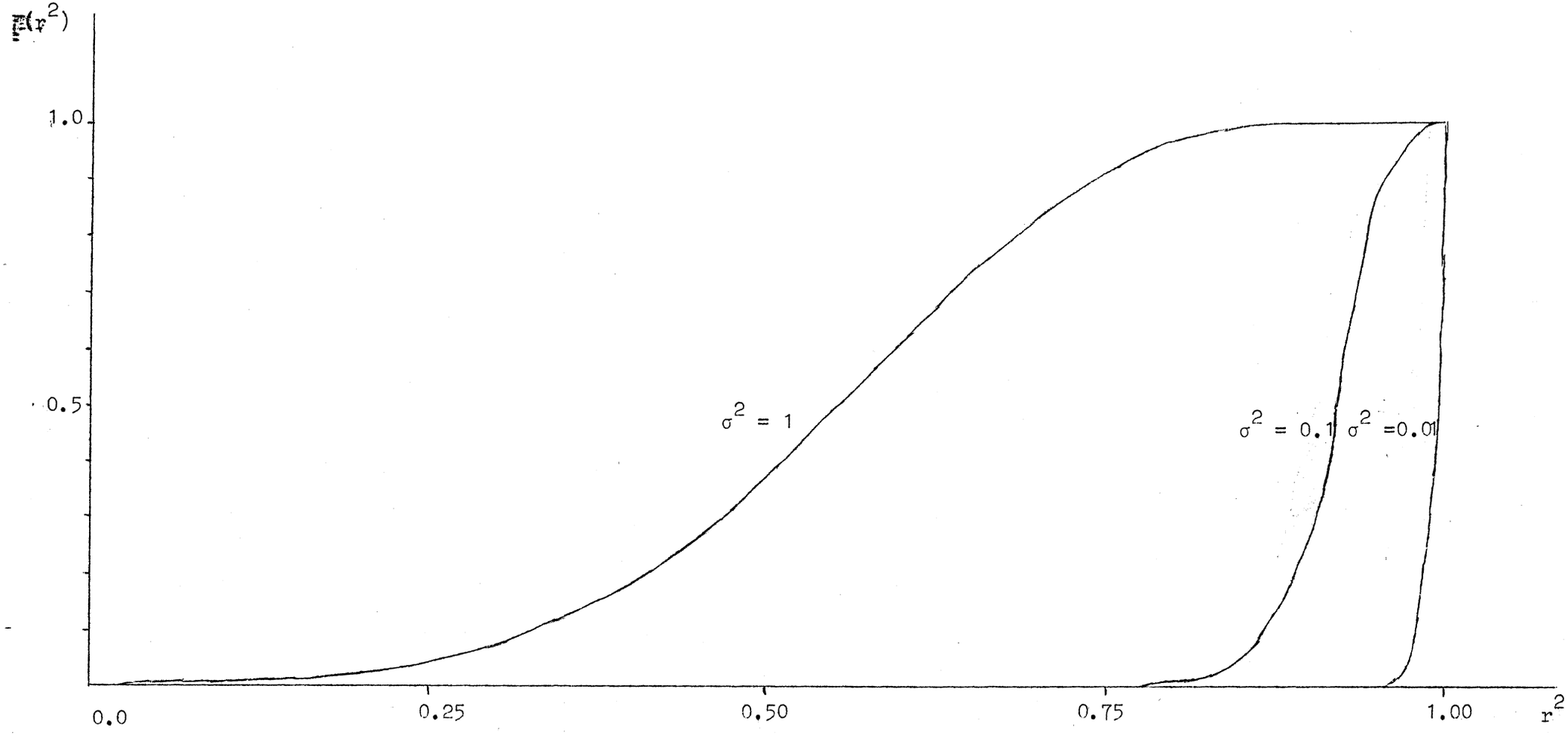
[4] Both matrices are given in the appendix.

Distribution Function of R^2



Textile Example ($\lambda = 3$)

Distribution Function of R^2



Artificial Example ($\lambda = 3$)

What are the causes of this striking difference? Whatever they may be, they can only influence the distribution of R^2 along two lines, via the characteristic roots λ_j of the matrix $(kE - M)$ or via the non-centrality parameters δ_j^2 for $j = 1, \ldots, n$. We first consider the characteristic roots λ_j and thus examine the matrix $E - M$. The matrix E is idempotent and has rank $n - 1$, for its trace is

$$(4.1) \qquad \operatorname{tr} E = \operatorname{tr} \left[I - \frac{1}{n}\iota\iota'\right] = \operatorname{tr} I - \operatorname{tr} \frac{1}{n}\iota\iota' = n - 1$$

Moreover

$$(4.2) \qquad E\iota = \left[I - \frac{1}{n}\iota\iota'\right]\iota = \iota - \frac{1}{n}\iota\iota'\iota = \iota - \iota = 0$$

which implies that the columns of E span the $n - 1$ dimensional orthogonal complement of ι. With respect to the matrix M, we have

$$(4.3) \qquad MX = [I - X(X'X)^{-1}X']X = X - X = 0$$

If the first column of X is equal to ι, which implies that we are dealing with a linear model containing a constant term,

$$(4.4) \qquad M\iota = 0$$

Thus the $n - \Lambda$ dimensional subspace spanned by the columns of M is orthogonal to ι. But this means that the space spanned by the columns of M is a subspace of the space spanned by the columns of E. Therefore

$$(4.5) \qquad EM = M \quad \text{and thus} \quad ME = M$$

because E is a projection matrix (E projects orthogonally along ι) and leaves all vectors in the orthogonal complement of ι unchanged. From (4.5), we can conclude that $E - M$ is idempotent, for

$$(4.6) \qquad (E - M)(E - M) = E^2 - EM - ME + M^2 = E - M - M + M = E - M$$

so its characteristic roots are either zero or one. The rank of $E - M$ is equal to its trace, equal to

$$(4.7) \qquad \operatorname{tr} (E - M) = \operatorname{tr} E - \operatorname{tr} M = n - 1 - (n - \Lambda) = \Lambda - 1$$

Therefore $\Lambda - 1$ of its roots are equal to 1; the others are equal to zero.

Let us now consider the characteristic roots of the matrix $kE - M$. This matrix has, in any case, one characteristic root equal to zero for

$$(4.8) \qquad (kE - M)\iota = 0$$

hence ι is an unscaled characteristic vector of $kE - M$ belonging to a zero root. To determine the other roots, consider the characteristic vectors p of M for which $Mp = p$. Furthermore we have $Ep = p$ because E projects any vector orthogonally on the orthogonal complement of ι while the p's are vectors in the latter space. We can then write

$$(4.9) \qquad (kE - M)p = kEp - Mp = kp - p = (k - 1)p$$

and hence, $n - \Lambda$ roots of $(kE - M)$ are equal to $k - 1$. To determine the remaining $\Lambda - 1$ characteristic roots of $(kE - M)$, we consider the characteristic vectors p of M (other then ι) for which $Mp = 0$. We can write

$$(4.10) \qquad (kE - M)p = (kE - M)Ep = kEp - MEp = kEp - Mp = kEp = kp$$

from which it follows that $\Lambda - 1$ roots of $(kE - M)$ are equal to k. Hence we have proved that the characteristic roots of $(kE - M)$ are independent of the X-matrix.

This important observation leads to the conclusion that the X-matrix influences the distribution of R^2 only through the non-centrality parameters δ_j^2 for $j = 1, \ldots, n$. We show, however, that only $\Lambda - 1$ of these non-centrality parameters really matter. Remember, that the vector of non-centrality parameters is equal to $P'X\beta/\sigma$ where P is a square matrix of order n consisting of the characteristic vectors of the matrix $(kE - M)$. We have already seen that $n - \Lambda$ of these lie in the space spanned by the columns of M, for they possess the property $Mp = p$, see (4.9). Since $MX = 0$, this implies that they lie in the space orthogonal to the space spanned by the columns of X.

Hence $n - \Lambda$ columns of the matrix P are orthogonal to the columns of X, which means that $n - \Lambda$ non-centrality parameters are equal to zero. In other words, at most, Λ non-centrality parameters differ from zero. The characteristic roots λ_j corresponding to the Λ non-central chi-square distributed variables z_j (see (3.7)) are the roots of $(kE - M)$ belonging to its characteristic vectors with the property $Mp = 0$. Thus $\Lambda - 1$ of them are equal to k and one is equal to zero (see (4.10)). This means that the X-matrix influences the distribution of R^2 through $\Lambda - 1$ non-centrality parameters.[5]

[5] Besides the number of observations of course.

In section 2, we have shown that the distribution of R^2 can be determined as follows

$$F(r^2) = P\left[\sum_{j=1}^{n} \lambda_j z_j^2 \leq 0\right] \tag{4.11}$$

If we change the value of r^2, we change the relevant roots λ_j for, apart from the zero root and the multiplicative constant σ^2, they are equal to $k = 1 - r^2$. This implies that the characteristic vectors of $kE - M$ change and so do the relevant non-centrality parameters δ_j^2. This fact is shown in Table 3 where the non-centrality parameters for both the textile example and the artificial example are given for different values of r^2. The variance of the disturbances is kept equal to 1.

TABLE 3

	Textile example		Artificial example	
	λ_j	δ_j^2	λ_j	δ_j^2
$r^2 = 0.1$				
	0.900	0.083	0.900	8.063
	0.900	0.002	0.900	5.108
	0.000	66.916	0.000	45.312
		$\sum_{j=1}^{3} \delta_j^2 = 67.001$		$\sum_{j=1}^{3} \delta_j^2 = 58.483$
$r^2 = 0.2$				
	0.800	0.083	0.800	7.163
	0.800	0.002	0.800	6.008
	0.000	66.916	0.000	45.312
		$\sum_{j=1}^{3} \delta_j^2 = 67.001$		$\sum_{j=1}^{3} \delta_j^2 = 58.483$
$r^2 = 0.4$				
	0.600	0.083	0.600	1.456
	0.600	0.002	0.600	11.715
	0.000	66.916	0.000	45.312
		$\sum_{j=1}^{3} \delta_j^2 = 67.001$		$\sum_{j=1}^{3} \delta_j^2 = 58.483$

Notice, that the non-centrality parameter corresponding to the zero root does not depend on the value of r^2. This is clear because its characteristic vector is identically equal to $1/\sqrt{n}$ and thus independent of r^2.

From the table we also learn that the sums of squares of the δ_j are equal to each other for any value of r^2. This is so because

$$\beta'X'P'PX\beta/\sigma^2 = \beta'X'X\beta/\sigma^2 \tag{4.12}$$

which means that the matrix of characteristic vectors cancels if we consider the sum of squares of the non-centrality parameters. We can now draw the interesting conclusion that knowledge of the individual δ_j^2 is of no importance if one is interested in the distribution of R^2. It is only the sum of squares of the relevant $\Lambda - 1$ non-centrality parameters that really matters. This is true for two reasons: (1) for a given value of r^2 all relevant λ_j are equal to each other except for the zero root. This root, however, causes no trouble because its corresponding non-centrality parameter is always the same and (2) the non-central chi-square distribution is additive, that is the sum of n independent non-central chi-square distributed variables is again non-central chi-square distributed with a non-centrality parameter equal to the sum of the individual δ_j^2's. This can easily be proved with the aid of its characteristic function

$$\varphi(t) = (1 - 2it)^{-\frac{1}{2}} \exp\left\{\frac{\delta^2 t}{1 - 2it}\right\} \tag{4.13}$$

Our statement follows from the fact that the characteristic function of a sum of independent random variables is equal to the product of the individual characteristic functions.

It seems therefore reasonable to use the sum of squares of the $\Lambda - 1$ relevant non-centrality parameters as a measure of skewness of the distribution of R^2. For a large sum of squares implies that the distribution of R^2 is concentrated at values close to one. For the practical research worker, this implies that the expression $\beta'X'X\beta$, or better an estimate of it, becomes important.

Finally, the difference between the textile and the artificial examples becomes clear if we consider the corresponding non-centrality parameters. In the artificial example, they are much larger. The distribution of R^2 is therefore shifted towards the right.

4.3. Changing Variables

In many cases, the "final" specification of an economic relationship is determined by introducing, one after another different explanatory variables. In practice, specification with the highest correlation coefficient is usually preferred. In this section, we discuss this strategy in more detail. Our procedure runs as follows. We start with a model with two explanatory variables. A constant term is added, thus $\Lambda = 3$. The explanatory variables will be denoted by X_1, X_2 and X_3.

The distribution of R^2 has been determined for different values of σ^2. The results are given in Table 4. Next, we introduce another explanatory variable X_4, and the distribution of R^2 is again computed. The results are given in Table 5.

Comparing Tables 4 and 5, it becomes clear how difficult it is to make a choice between these two linear models on the basis of the correlation coefficient. In fact , correlation coefficients belonging to different linear models can hardly be compared, for each coefficient should be interpreted in the light of its own probability distribution. However, it is precisely this distribution which depends on the specific explanatory variables used in the linear model. From Table 5, we see that the influence of the introduction of X_4 cannot be ignored. The distribution of R^2 changes considerably. In the case of three explanatory variables, 82.97 percent of the total probability mass lies in the interval [0, 0.7] while, in the case of four explanatory variables, the percentage is much lower, 37.9. This means that, on purely technical grounds, a much larger correlation coefficient should be expected. In Table 8, we see the effect of the introduction of an additional explanatory variable X_j; the distribution of R^2 is shifted even more towards the right.

In the Tables 6 and 7, we have kept the number of explanatory variables equal to 4, but the models differ from each other by one explanatory variable. Also from these Tables we see how sensitive the distribution of R^2 is to the X-matrix. Here again we see how difficult it is to evaluate different values of R^2 without knowing the underlying distribution.

TABLE 4. THE DISTRIBUTION FUNCTIONS OF R^2 FOR THE EXPLANATORY VARIABLES X_1, X_2, X_3

R^2	$\sigma^2 = 9$	$\sigma^2 = 16$	$\sigma^2 = 25$
0.10	0.0003	0.0098	0.0463
0.20	0.0039	0.0565	0.1718
0.30	0.0241	0.1741	0.3742
0.40	0.0954	0.3760	0.6091
0.50	0.2667	0.6251	0.8115
0.60	0.5461	0.8410	0.9370
0.70	0.8297	0.9614	0.9880
0.80	0.9760	0.9966	0.9992
0.90	0.9997	0.9999	0.9999
1.00	1.0000	1.0000	1.0000
$\sum_{j=1}^{2} \delta_j^2$	22.1547	12.4621	7.9757

TABLE 5. THE DISTRIBUTION FUNCTIONS OF R^2 FOR THE EXPLANATORY VARIABLES X_1, X_2, X_3, X_4

R^2	$\sigma^2 = 9$	$\sigma^2 = 16$	$\sigma^2 = 25$
0.10	0.0000	0.0001	0.0025
0.20	0.0000	0.0021	0.0211
0.30	0.0002	0.0147	0.0858
0.40	0.0025	0.0656	0.2321
0.50	0.0199	0.2050	0.4656
0.60	0.1082	0.4649	0.7269
0.70	0.3795	0.7716	0.9164
0.80	0.7959	0.9607	0.9901
0.90	0.9924	0.9993	0.9999
1.00	1.0000	1.0000	1.0000
$\sum_{j=1}^{3} \delta_j^2$	39.5398	22.2408	14.2341

TABLE 6. THE DISTRIBUTION FUNCTIONS OF R^2 FOR THE EXPLANATORY VARIABLES X_1, X_2, X_3, X_5

R^2	$\sigma^2 = 9$	$\sigma^2 = 16$	$\sigma^2 = 25$
0.10	0.0000	0.0000	0.0004
0.20	0.0000	0.0001	0.0055
0.30	0.0000	0.0025	0.0312
0.40	0.0001	0.0175	0.1135
0.50	0.0025	0.0821	0.2950
0.60	0.0264	0.2684	0.5710
0.70	0.1727	0.6020	0.8389
0.80	0.6052	0.9077	0.9761
0.90	0.9756	0.9977	0.9996
1.00	1.0000	1.0000	1.0000
$\sum_{j=1}^{3} \delta_j^2$	52.6035	29.5895	18.9373

TABLE 7. THE DISTRIBUTION FUNCTIONS OF R^2 FOR THE EXPLANATORY VARIABLES X_1, X_2, X_4, X_5

R^2	$\sigma^2 = 9$	$\sigma^2 = 16$	$\sigma^2 = 25$
0.10	0.0000	0.0002	0.0032
0.20	0.0000	0.0029	0.0258
0.30	0.0003	0.0191	0.0955
0.40	0.0037	0.0794	0.2570
0.50	0.0269	0.2330	0.4956
0.60	0.1316	0.5003	0.7499
0.70	0.4209	0.7955	0.9260
0.80	0.8216	0.9665	0.9915
0.90	0.9939	0.9994	0.9999
1.00	1.0000	1.0000	1.0000
$\sum_{j=1}^{3} \delta_j^2$	37.5362	21.1141	13.5130

TABLE 8. THE DISTRIBUTION FUNCTION OF R^2 FOR THE EXPLANATORY VARIABLES X_1, X_2, X_3, X_4 AND X_5

R^2	$\sigma^2 = 9$	$\sigma^2 = 16$	$\sigma^2 = 25$
0.10	0.0000	0.0000	0.0000
0.20	0.0000	0.0000	0.0006
0.30	0.0000	0.0002	0.0060
0.40	0.0000	0.0023	0.0332
0.50	0.0001	0.0186	0.1263
0.60	0.0032	0.1004	0.3433
0.70	0.0458	0.3562	0.6672
0.80	0.3384	0.7701	0.9263
0.90	0.9132	0.9893	0.9981
1.00	1.0000	1.0000	1.0000
$\sum_{j=1}^{4} \delta_j^2$	66.0293	37.1415	23.7706

5. SOME CONCLUDING REMARKS

We now summarize the results:

(1) The distribution of the correlation coefficient depends on the matrix X of explanatory variables. Its probability distribution is certainly not robust with respect to changes in this matrix.

(2) The X-matrix influences the distribution of R^2 only through $\Lambda - 1$ non-centrality parameters δ_j^2.

(3) The variance σ^2 influences the distribution of R^2 also via the $\Lambda - 1$ non-centrality parameters, for σ^2 appears in the denominator of the non-centrality parameters $\delta_j = P'X\beta/\sigma$ and also via the corresponding λ_j.

(4) It is dangerous to compare correlation coefficients belonging to distinct models because, for a given value of σ^2, this distributions depend on the X-matrices. For a given value of σ^2 an increase of R^2 may happen for purely technical reasons.

(5) The correlation coefficient should be interpreted in the light of its own distribution. This, however, can only be done if we know the value of σ^2. For this purpose, we cannot use an estimate of σ^2 based on the least-squares disturbances, because the choice of the X-matrix implies an estimate of σ^2. The only way out seems to start with an a priori value for σ^2. Given this value and a particular X-matrix; the distribution of R^2 can be determined. The following strategy to check the reality of the model can then be used. If the value of R^2 found in practice is acceptable in the light of this distribution we accept the model, if it is highly unprobable, we reject the model and look for another one.

In a future paper the authors will discuss the influence of auto-correlation on the distribution of R^2. Moreover, they will examine the question of how the sum of squares of δ_j^2 for $j = 1, \ldots, \Lambda - 1$ can be used to get a quick information about the distribution of R^2.

REFERENCES

[1] Imhof, J.P. (1961). "Computing the Distribution of Quadratic Forms in Normal Variables". Biometrika , Vol. 48, 3 and 4, pp. 419-426.

A P P E N D I X

The X-matrix of the textile example is:

1	1.9851	2.0043
1	1.9917	2.0004
1	2.0000	2.0000
1	2.0208	1.9571
1	2.0208	1.9370
1	2.0394	1.9528
1	2.0445	1.9571
1	2.0504	1.9180
1	2.0386	1.8457
1	2.0224	1.8156
1	2.0073	1.7875
1	1.9796	1.7959
1	1.9841	1.8035
1	1.9895	1.7210
1	2.0103	1.7760

and the coefficients are:

$\beta_1 = 1.3739$

$\beta_2 = 1.1432$

$\beta_3 = -0.8289$

The X-matrix of the artificial example is:

1	1.4794	1.8776
1	1.8415	1.5403
1	1.9975	1.0707
1	1.9093	0.5839
1	1.5985	0.1989
1	1.1411	0.0100
1	0.6492	0.0635
1	0.2432	0.3464
1	0.0225	0.7892
1	0.0411	1.2837
1	0.2945	1.7087
1	0.7206	1.9602
1	1.2151	1.9766
1	1.6570	1.7539
1	1.9380	1.3466

and the coefficients are:

$\beta_1 = 1.3739$

$\beta_2 = 1.1432$

$\beta_3 = -0.8289$