

## MEAN AND VARIANCE OF $R^2$ IN SMALL AND MODERATE SAMPLES\*

J.S. CRAMER

*University of Amsterdam, 1011 NH Amsterdam, The Netherlands*

Received March 1986, final version received September 1986

We derive and use easily computable expressions for the mean and variance of  $R^2$  in the standard linear regression model with fixed regressors. In respect to its probability limit  $R^2$  is seriously biased upward in small samples; the 'adjusted'  $\bar{R}^2$  does much better. But at sample sizes where these distinctions matter both measures are thoroughly unreliable because of their large dispersion.  $R^2$  should not be quoted for samples of less than fifty observations.

### 1. Introduction

Ordinary least squares (OLS) estimation of linear regression equations is still an accepted tool of analysis among economists. The reported results invariably include  $R^2$  or the 'adjusted'  $\bar{R}^2$ . These measures of goodness of fit have a fatal attraction. Although it is generally conceded among insiders that they do not mean a thing, high values are still a source of pride and satisfaction to their authors, however hard they may try to conceal these feelings.

To put these sample statistics in proper perspective we shall derive their means and variances for various sample sizes under the standard assumptions of econometric theory. This means that the regressor variables are regarded as given, non-random constants. In this respect the model differs from the classical treatment of correlation in the setting of a multivariate Normal distribution, and the results differ too. The mean of  $R^2$  converges to its probability limit from above, and in this sense it has an upward bias which can be substantial in small samples. In this respect  $\bar{R}^2$  is superior. The standard errors show however that for sample sizes of up to 40 or 50 either measure is a very unreliable statistic.

\*I have benefited from the corrections and improvements suggested by several referees, Norman Draper, Joyce Meijering, and V. Srivastava. I owe particular debts to Roald Ramer and Jan de Leeuw for their help with the algebra and bibliography of section 4.

## 2. The moments of $R^2$ in the standard model

We consider the standard linear regression model with Normal disturbances as given in any textbook of econometrics, but we employ a slightly non-standard presentation and notation. We write

$$y = c\gamma + X\beta + \varepsilon, \quad (1)$$

with  $y$  a  $(m \times 1)$  vector of observed values of the dependent variable and  $\varepsilon$  a vector of  $m$  independent  $N(0, \sigma^2)$  disturbances. On the right  $c$  is a unit vector,  $\gamma$  the intercept parameter, and  $X$  a matrix of  $(k-1)$  regressor variables, which have all been measured as deviations from their sample means. This last property of  $X$  simplifies the notation in the sequel, like the use of  $m$  instead of  $n$  for the actual sample size. Note that we have merely reparametrized the systematic, non-random part of the right-hand side, without touching the definition of  $y$  and  $\varepsilon$ ; we have *not* taken deviations from the mean for the dependent variable, and the elements of  $\varepsilon$  are still stochastically independent.

With ordinary least squares, the estimate of  $\gamma$  is

$$\bar{y} = \frac{1}{m} c'y, \quad (2)$$

which is the sample mean of  $y$ , while the estimate of  $\beta$  is

$$b = (X'X)^{-1} X'y. \quad (3)$$

Upon defining the residual vector  $e$  as

$$e = y - c\bar{y} - Xb, \quad (4)$$

we have the identity

$$y'y = m\bar{y}^2 + b'X'Xb + e'e,$$

or

$$(y'y - m\bar{y}^2) = b'X'Xb + e'e. \quad (5)$$

This is the familiar decomposition of the sum of squares of  $y$ , on the left, into a systematic and a residual component. Their relative size determines the familiar measure of goodness of fit

$$R^2 = 1 - \frac{e'e}{y'y - m\bar{y}^2} = \frac{b'X'Xb}{b'X'Xb + e'e}. \quad (6)$$

We derive the density function of  $R^2$  and then its moments for the general case with  $\beta \neq 0$ . First consider the transformation

$$G = \frac{R^2}{1 - R^2} = \frac{b'X'Xb/\sigma^2}{e'e/\sigma^2}. \quad (7)$$

$G$  (which is halfway towards  $F$ ) is the ratio of two independent chi-square variates. The numerator has a non-central chi-square distribution,

$$\frac{b'X'Xb}{\sigma^2} \sim \chi^2(\lambda, k-1), \quad (8)$$

with  $(k-1)$  degrees of freedom and non-centrality parameters

$$\lambda = \frac{\beta'X'X\beta}{\sigma^2}. \quad (9)$$

For the denominator we have of course

$$\frac{e'e}{\sigma^2} \sim \chi^2(0, m-k). \quad (10)$$

The density function of  $G$  can be found, for example, in Johnson and Kotz (1970, II, p. 191).<sup>1</sup> Upon introducing the transformation (7) and relabelling the parameters we obtain the density of  $R^2$ , with argument  $r$ , from the transformation theorem. This yields

$$f(r) = \sum_{j=0}^{\infty} w(j) \frac{1}{B(u+j, v-u)} r^{u+j-1} (1-r)^{v-u-1}, \quad (11a)$$

with

$$w(j) = \frac{e^{-\frac{1}{2}\lambda} \left(\frac{1}{2}\lambda\right)^j}{j!}, \quad (11b)$$

$$u = \frac{1}{2}(k-1), \quad (11c)$$

$$v = \frac{1}{2}(m-1). \quad (11d)$$

<sup>1</sup>A misprint in the first line of (5) gives  $\nu_1$  where  $\nu_2$  is intended.

Making use of the properties of the Beta function [see Abramovitz and Stegun (1964, pp. 256–258)] we obtain the moments of  $R^2$  as

$$E(R^2) = \sum_{j=0}^{\infty} w(j) \frac{u+j}{v+j}, \quad (12a)$$

$$E(R^2)^2 = \sum_{j=0}^{\infty} w(j) \frac{u+j}{v+j} \frac{u+j+1}{v+j+1}, \quad (12b)$$

and so forth. Their dependence on the three parameters  $\lambda$ ,  $k$  and  $m$  is clear. We shall shortly see that they are quite easy to compute.

### 3. $\lambda$ and the probability limit of $R^2$

Eq. (9) defines the parameter  $\lambda$  as the ratio of the systematic variation  $\beta'X'X\beta$  to the disturbance variance  $\sigma^2$ . These magnitudes differ by a factor  $m$ , as can be seen by taking expectations on both sides of eq. (5),

$$y'y - m\bar{y}^2 = b'X'Xb + e'e.$$

Neglecting the loss of degrees of freedom among the residuals, we find the expected sum of squares of  $Y$  as

$$SSY = \beta'X'X\beta + m\sigma^2. \quad (13)$$

By analogy to the passage from (5) to (6) this naturally suggests

$$\phi = \frac{\beta'X'X\beta}{\beta'X'X\beta + m\sigma^2}, \quad (14)$$

for a measure of the quality of fit as determined by the underlying conditions of the observations, itself free from sampling variation. This magnitude is commensurate with  $R^2$ , and it is related to  $\lambda$  of (9) by

$$\phi = \frac{\lambda}{\lambda + m}, \quad \lambda = m \frac{\phi}{1 - \phi}. \quad (15)$$

This parametrization is standard in earlier analyses of  $R^2$  with fixed, non-random  $X$ , as opposed to the case of a joint multivariate Normal distribution of the elements of  $y$  and the regressor variables of  $X$ . Barten (1962) defined  $\phi$  without further ado as the 'parent multiple correlation coefficient' that is estimated by  $R^2$ , and Schönfeld (1969, p. 71) equally relies

entirely on intuitive appeal when he labels  $\phi$  the 'theoretical measure of fit'. Press and Zellner (1978) take the parameter from Barten as they lay the basis for a Bayesian analysis.

We shall follow the example of Koerts and Abrahamse (1970) and derive  $\phi$  as the probability limit of  $R^2$ .<sup>2</sup> To do so we introduce sample size  $n$  as a variable which has the value  $m$  for the sample actually observed. We also rewrite (6) in self-evident notation as

$$R_n^2 = \frac{b_n'(n^{-1}X_n'X_n)b_n}{b_n'(n^{-1}X_n'X_n)b_n + n^{-1}e_n'e_n}. \quad (16)$$

In passing to the limit for  $n \rightarrow \infty$ , the main difficulty is the behaviour of  $X_n$  since this consists of non-random constants. We resort to the device of Hotelling (1940) to treat the regressors as 'constant in repeated samples'. This means that the virtual sample size  $n$  is given by

$$n = pm, \quad (17)$$

with integer  $p$ , and that the matrix  $X_n$  consists of  $p$  replications of  $X = X_m$ , stacked on top of one another. We vary  $n$  by varying  $p$ , and thus obtain, just like Theil (1971, p. 363),

$$\text{plim}_{n \rightarrow \infty} (n^{-1}X_n'X_n) = \lim_{p \rightarrow \infty} \left( p^{-1} \sum_{j=1}^p m^{-1}X_m'X_m \right) = m^{-1}X'X. \quad (18)$$

With  $X_n$  behaving in this fashion and i.i.d. disturbances (as assumed) the OLS estimate is consistent,

$$\text{plim}_{n \rightarrow \infty} b_n = \beta, \quad (19)$$

and

$$\text{plim}_{n \rightarrow \infty} n^{-1}e_n'e_n = \sigma^2. \quad (20)$$

Upon substitution of these three probability limits into (15) we obtain, by (14),

$$\text{plim}_{n \rightarrow \infty} R_n^2 = \frac{\beta'(m^{-1}X'X)\beta}{\beta'(m^{-1}X'X)\beta + \sigma^2} = \phi, \quad (21)$$

which is the desired result.

<sup>2</sup>Some of these authors pursue the same questions as we do. Barten derives an approximate expression for the bias of  $R^2$  relative to  $\phi$ , and suggests corrections, but he does not examine dispersion. Koerts and Abrahamse establish the distribution of  $R^2$  for given  $\sigma^2$ ,  $\beta$  and  $X$ , and show that this is very sensitive to changes in  $X$ . The distribution is determined numerically, and it must be computed anew for each new matrix  $X$ .

This provides a direct link of the parameter  $\phi$  with  $R^2$  in the model under consideration. The identification with a probability limit is particularly appropriate as we shall examine the behaviour of  $R^2$  at different sample sizes.

#### 4. The mean and standard deviation of $R^2$

We return to eq. (12) for the first two moments of  $R^2$ . While (12a) may be developed a little further (as we shall presently see), we can evaluate both expressions to any desired degree of accuracy by summing the first  $J$  terms of the infinite series concerned. When we write the moments as given in (12) as

$$\sum_{j=0}^{\infty} w(j)z(j),$$

the discrepancy involved in taking the first  $J$  terms only is

$$\delta = \sum_{j=J+1}^{\infty} w(j)z(j). \quad (22)$$

For all moments the  $z(j)$  are positive and tend from below to 1 as  $j \rightarrow \infty$ , and  $w(j)$  is a Poisson density which sums to 1. Clearly, then,

$$0 < \delta < 1 - \sum_{j=0}^J w(j). \quad (23)$$

It requires no great programming skill to continue summing the  $w(j)$ ,  $z(j)$  for given parameter values until the right-hand side of (23) reduces  $\delta$  to the desired level of accuracy. In this sense (12) provides easily computable expressions for the moments of  $R^2$ .

In the event we have set  $\delta$  at  $10^{-8}$  in computing the first two moments for various values of  $\phi$ ,  $m$  and  $k$ . The mean is overly accurate, and the standard deviation derived from the two moments is correct to three decimal places. The results are given in table 1, and illustrated in figs. 1 and 2.

Fig. 1 shows that  $E(R^2)$  converges rather quickly to  $\phi$  from above. Some simulations, not further reported here, suggest that this also holds for the median and the mode.  $R^2$  thus has a definite upwards bias which is however rapidly reduced as the sample size increases. Very roughly the bias is about 0.03 or less with twenty observations when we have one regressor ( $k = 2$ ), or

Table 1  
Mean and standard deviation of  $R^2$  for selected values of  $k$ ,  $\phi$ , and of  $m$ .

$\phi =$	Mean									Standard deviation								
	$k = 2$						$k = 3$			$k = 2$						$k = 3$		
	0.9	0.667	0.5	0.333	0.9	0.667	0.5	0.333	0.9	0.667	0.5	0.333	0.9	0.667	0.5	0.333		
$m = 5$	0.936	0.760	0.620	0.485	0.957	0.840	0.747	0.657	0.052	0.183	0.254	0.287	0.042	0.152	0.217	0.260		
6	0.930	0.742	0.597	0.454	0.947	0.807	0.697	0.590	0.049	0.171	0.237	0.267	0.042	0.150	0.213	0.251		
7	0.925	0.730	0.581	0.433	0.940	0.784	0.665	0.547	0.047	0.160	0.222	0.251	0.042	0.145	0.204	0.248		
8	0.922	0.722	0.569	0.419	0.935	0.768	0.641	0.516	0.045	0.151	0.209	0.237	0.041	0.139	0.196	0.229		
9	0.920	0.715	0.561	0.408	0.931	0.756	0.624	0.493	0.043	0.143	0.198	0.225	0.039	0.133	0.188	0.219		
10	0.918	0.710	0.554	0.400	0.928	0.746	0.610	0.475	0.041	0.136	0.189	0.214	0.038	0.128	0.180	0.210		
20	0.908	0.688	0.526	0.365	0.914	0.705	0.552	0.400	0.030	0.098	0.136	0.155	0.029	0.096	0.133	0.153		
30	0.906	0.681	0.517	0.354	0.909	0.692	0.534	0.377	0.025	0.080	0.111	0.127	0.025	0.079	0.110	0.126		
40	0.904	0.677	0.513	0.349	0.907	0.686	0.526	0.366	0.022	0.070	0.096	0.110	0.022	0.069	0.096	0.110		
50	0.903	0.675	0.510	0.345	0.905	0.682	0.520	0.359	0.020	0.062	0.086	0.099	0.019	0.062	0.086	0.098		
100	0.902	0.671	0.505	0.339	0.903	0.674	0.510	0.346	0.014	0.044	0.061	0.070	0.014	0.044	0.061	0.070		
150	0.901	0.670	0.503	0.337	0.902	0.672	0.507	0.342	0.011	0.036	0.050	0.057	0.011	0.036	0.050	0.057		
200	0.901	0.668	0.502	0.336	0.901	0.670	0.504	0.338	0.009	0.028	0.039	0.044	0.009	0.028	0.039	0.044		

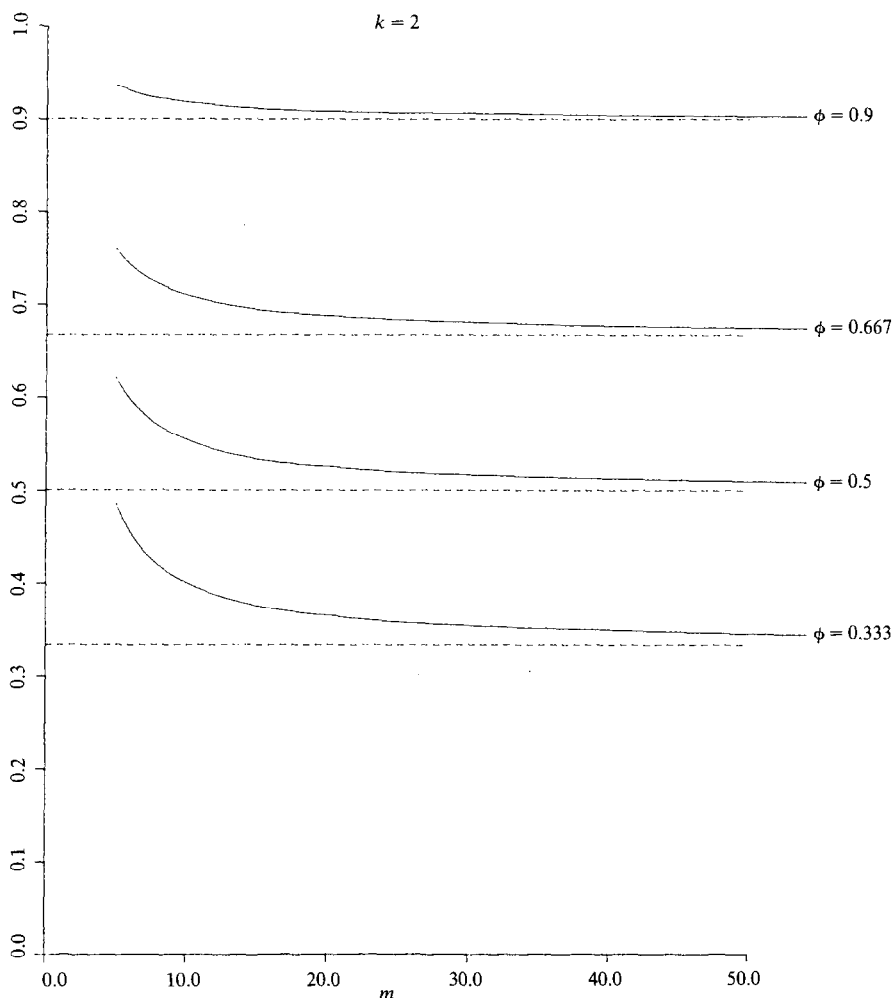


Fig. 1a. Expected value of  $R^2$  as a function of  $m$  for  $k = 2$  and selected values of  $\phi$ .

with thirty to forty observations with two regressors. But this is a rough indication only, as the exact values vary with  $\phi$ .

The reason for this brief dismissal of the bias is that it is completely swamped by the dispersion of  $R^2$ : whenever the bias is at all noticeable, the standard error of  $R^2$  is several times as large. Fig. 2 shows how the standard error varies with  $\phi$  and  $m$  and, to a much lesser extent, with  $k$ ; the effect of  $\phi$  is particularly strong. For a standard error of  $R^2$  of 0.03 or less we must have at least twenty observations if  $\phi = 0.9$ . Such high values of the true correlation



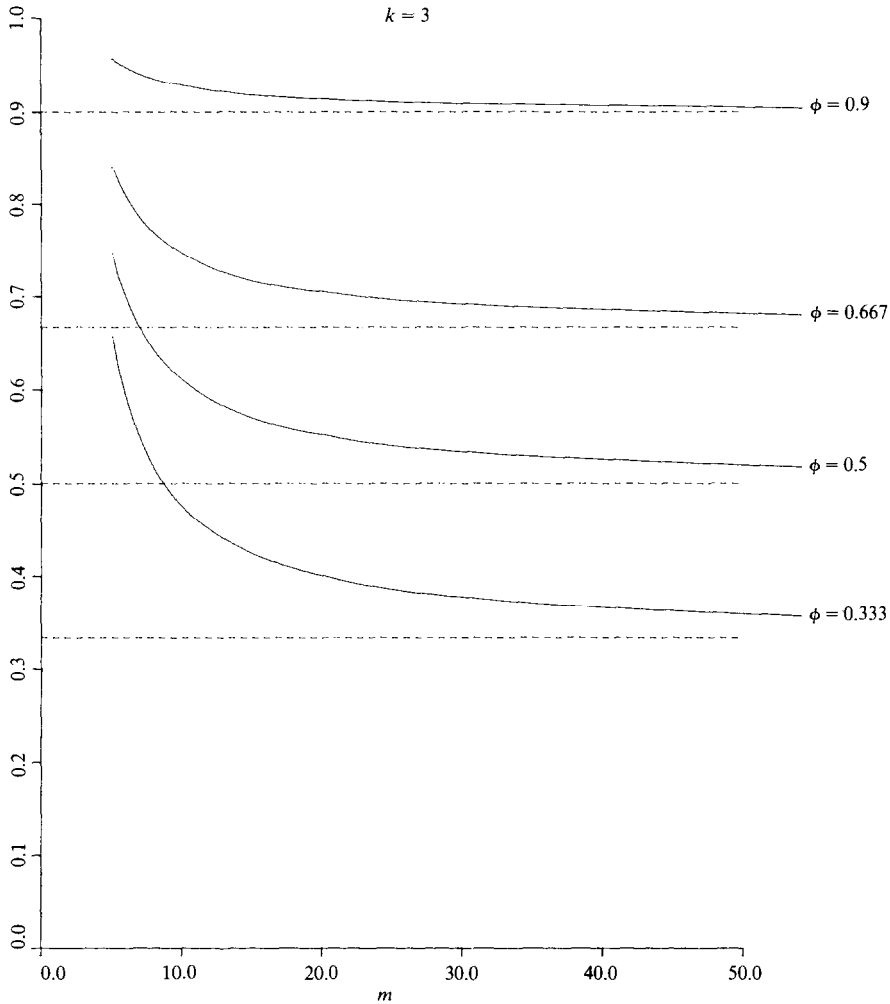


Fig. 1b. Expected value of  $R^2$  as a function of  $m$  for  $k = 3$  and selected values of  $\phi$ .

coefficient are probably exceptional (as opposed to *sample*  $R^2$  of 0.9); but with  $\phi$  at 0.667, which is still quite respectable, nearly two hundred observations are needed to reduce the standard error to 0.03. It is this dependence of the dispersion of the sample  $R^2$  on the unknown  $\phi$  which renders any judgment of accuracy so hazardous. The relationship is further illustrated in table 2, which shows how many observations are needed at various  $\phi$  to reduce the standard error of  $R^2$  to certain given levels.

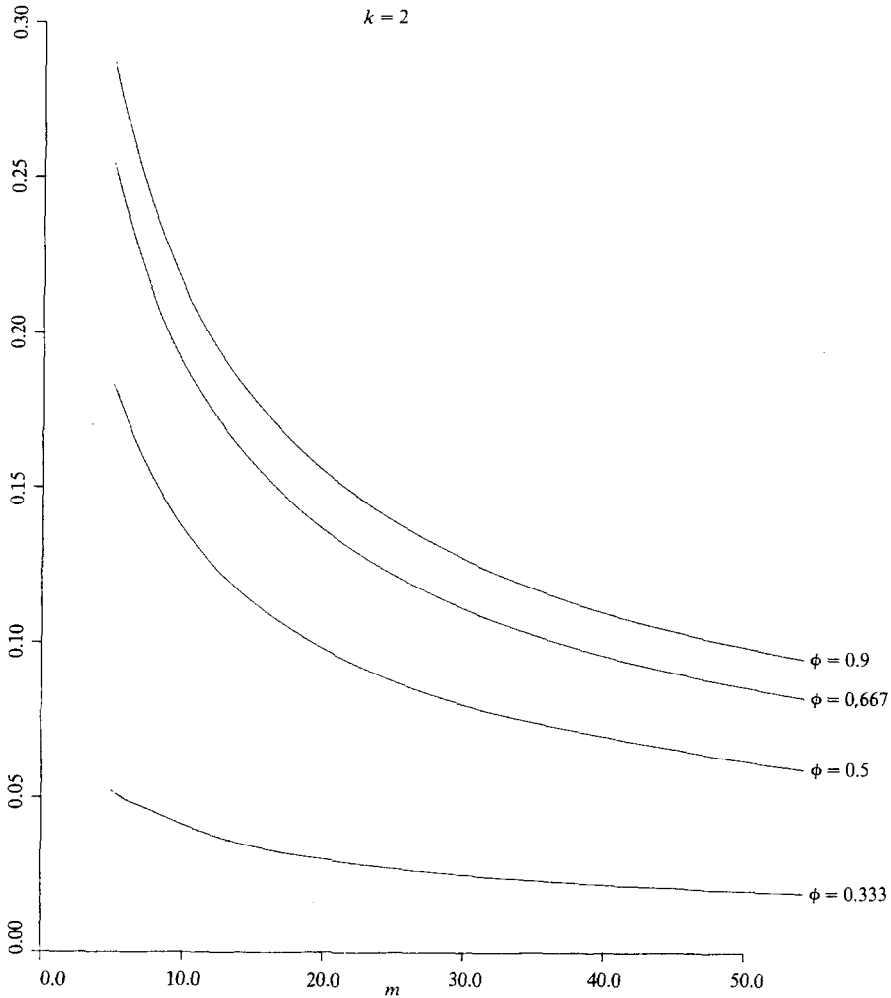


Fig. 2a. Standard deviation of  $R^2$  as a function of  $m$  for  $k = 2$  and selected values of  $\phi$ .

To sum up,  $R^2$  has an upward bias which can be substantial in small samples, but it is anyhow very unreliable, even at moderate sample sizes, because of its dispersion. With less than fifty observations or so there is little point in quoting  $R^2$  at all, and once we are beyond such numbers the bias has virtually disappeared. The bias issue is a red herring.

### 5. The adjusted multiple correlation coefficient $\bar{R}^2$

The intuitive explanation of the upward bias of  $R^2$  is that OLS treats it as the sample maximand, and the reason why this bias occurs in small samples is

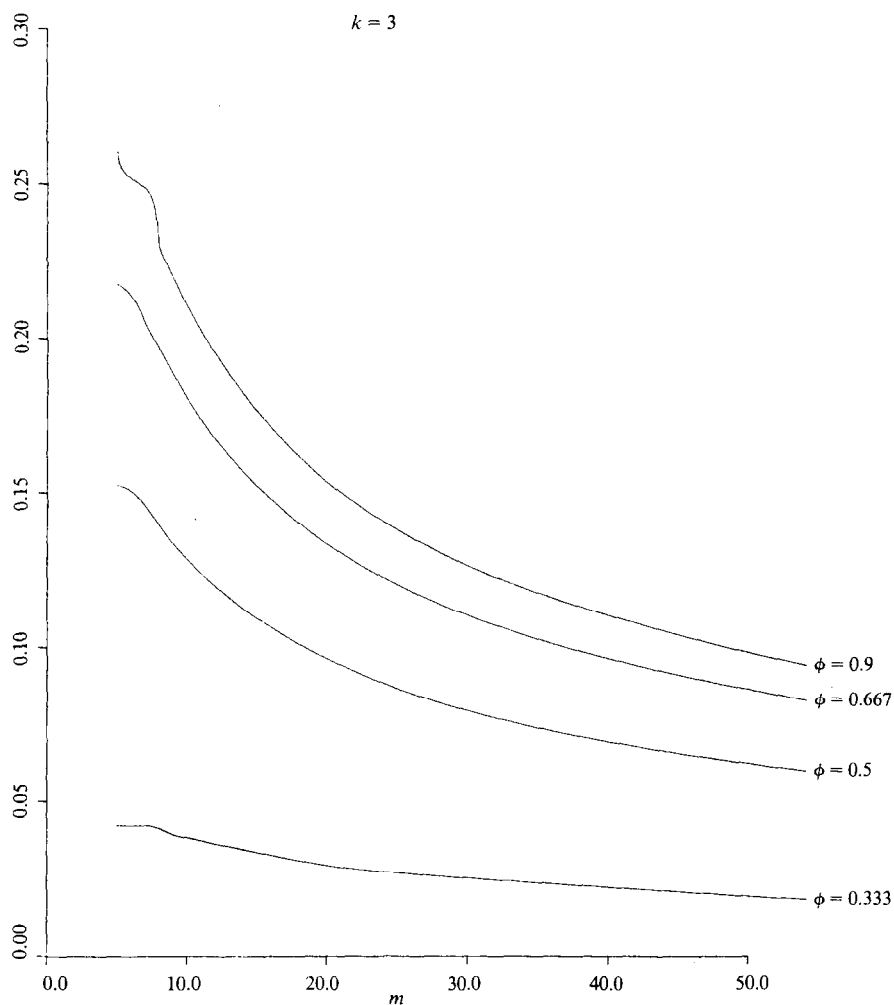


Fig. 2b. Standard deviation of  $R^2$  as a function of  $m$  for  $k = 3$  and selected values of  $\phi$ .

that  $R^2$  does not allow for the loss of degrees of freedom through estimation. This argument justifies the prevalent custom of 'adjusting'  $R^2$  as in

$$\bar{R}^2 = 1 - \frac{e'e/(m-k)}{(y'y - m\bar{y}^2)/(m-1)} = 1 - \frac{m-1}{m-k}(1-R^2), \quad (24)$$

or

$$\bar{R}^2 = (1+h)R^2 - h, \quad (25a)$$

Table 2  
Minimum sample size which reduces the standard error of  $R^2$  below a certain level  $\alpha$ .

$\alpha =$	$k = 2$			$k = 3$		
	0.03	0.05	0.10	0.03	0.05	0.10
$\phi = 0.30$	555	200	50	555	200	50
0.40	512	184	46	512	184	46
0.50	417	150	38	416	149	37
0.60	301	108	27	298	107	26
0.70	182	65	16	181	64	15
0.75	130	47	11	129	46	10
0.80	85	30	7	84	29	— <sup>a</sup>
0.85	48	17	— <sup>a</sup>	47	15	— <sup>a</sup>
0.90	21	6	— <sup>a</sup>	19	— <sup>a</sup>	— <sup>a</sup>
0.95	— <sup>a</sup>	— <sup>a</sup>	— <sup>a</sup>	— <sup>a</sup>	— <sup>a</sup>	— <sup>a</sup>

<sup>a</sup>All sample sizes  $\geq k + 1$  reduce the standard error below  $\alpha$ .

with

$$h = \frac{k - 1}{m - k}. \quad (25b)$$

As the first expression in (24) shows, the sums of squares in the first definition of  $R^2$  in (6) are 'corrected for degrees of freedom'. This adjustment is common usage among economists, probably because of their unique habit of submitting quite small samples to regression analysis. The adjustment is recommended in most econometric textbooks, all the way back to Ezekiel (1930a), though generally without much theoretical justification and without a source reference. The issue of correcting  $R^2$  in some way dates from the 1920's, for Ezekiel (1930b) can quote three slightly different definitions of  $h$  of (25a) from that decade. The surviving definition (25b) is due to Fisher (1924), who justifies it with the standard argument about sums of squares around the mean that is implicit in the first expression of (24).

More precise arguments in support of the adjustment (25) can be advanced. The adjusted  $\bar{R}^2$  does not suffer from the defect of  $R^2$  that is automatically increases with the addition of new regressors; we shall show the  $E(\bar{R}^2)$  is indeed independent of the number of regressors.

To begin with we must develop  $E(R^2)$  from (11) a little further, rewriting it as

$$\begin{aligned}
 E(R^2) &= e^{-\frac{1}{2}\lambda} \sum_{j=0}^{\infty} \frac{\left(\frac{1}{2}\lambda\right)^j}{j!} \frac{u+j}{v+j} \\
 &= e^{-\frac{1}{2}\lambda} \left\{ u \sum_j \frac{\left(\frac{1}{2}\lambda\right)^j}{j!(v+j)} + \frac{1}{2}\lambda \sum_j \frac{\left(\frac{1}{2}\lambda\right)^j}{j!(v+1+j)} \right\}. \quad (26)
 \end{aligned}$$

Upon consulting Slater (1964, p. 504, 13.1.2) we find that the summations within brackets are a special form of the Kummer function  $M(a, b, \mu)$ , namely  $a^{-1}M(a, a+1, \mu)$ . We introduce the notation

$$g(\mu, a) a^{-1} M(a, a+1, \mu) = \sum_{j=0}^{\infty} \frac{\mu^j}{j!(a+j)}, \quad (27)$$

and observe from Slater (1964, p. 505, 13.2.1) that

$$g(\mu, a) = \int_0^1 e^{\mu t} t^{a-1} dt. \quad (28)$$

Integrating the right-hand side by parts we find

$$\mu g(\mu, a+1) = e^{\mu} - ag(\mu, a). \quad (29)$$

We now make the appropriate substitutions of these results in (26) and obtain

$$E(R^2) = 1 - (v-u)e^{-\frac{1}{2}\lambda} g\left(\frac{1}{2}\lambda, v\right),$$

or, by (11c) and (11d),

$$E(R^2) = 1 - (m-k)_{\frac{1}{2}} e^{-\frac{1}{2}\lambda} g\left\{\frac{1}{2}\lambda, \left(\frac{1}{2}m-1\right)\right\}. \quad (30)$$

By (24), then,

$$E(\bar{R}^2) = 1 - (m-1)_{\frac{1}{2}} e^{-\frac{1}{2}\lambda} g\left\{\frac{1}{2}\lambda, \left(\frac{1}{2}m-1\right)\right\}, \quad (31)$$

and this expression depends only on  $m$  and on  $\phi$  (via  $\lambda$ ) but not on  $k$ :  $E(\bar{R}^2)$  is independent of the number of regressors.

We finally note, with more relevance to our original purpose, that the adjustment very largely removes the upward bias of  $R^2$ . By (25) we have

$$E(\bar{R}^2) = (1+h)E(R^2) - h, \quad (32)$$

and if this operation is applied to the entries of table 1 it will be seen that the bias virtually disappears; for low values of  $\phi$ , a slight downward bias occurs instead. But the dispersion remains, and is even increased, since

$$\text{s.d.}(\bar{R}^2) = (1+h)\text{s.d.}(R^2) \quad (33)$$

while  $h$  is positive and sizeable in small samples. In smallish samples  $\bar{R}^2$ , though unbiased, is even more unreliable than  $R^2$ .

## References

- Barten, A.P., 1962, Note on the unbiased estimation of the squared multiple correlation coefficient, *Statistica Neerlandica* 16, 151–163.
- Davis, P.J., 1964, Gamma functions and related functions, in: M. Abramovitz and I.A. Stegun, eds., *Handbook of mathematical functions* (Dover, New York) 253–293.
- Ezekiel, M., 1930a, *Methods of correlation analysis* (Wiley, New York).
- Ezekiel, M., 1930b, The sampling variability of linear and curvilinear regressions, *Annals of Mathematical Statistics* 1, 275–300.
- Fisher, R.A., 1924, The influence of rainfall on the yield of wheat at Rothamsted, *Philosophical Transactions of the Royal Society of London B* 213, 89–142.
- Hotelling, H., 1940, The selection of variates for use in prediction, *Annals of Mathematical Statistics* 11, 271–283.
- Johnson, N.L. and S. Kotz, 1970, *Distributions in statistics: Continuous univariate distributions*, Vol. 2 (Wiley, New York).
- Koerts, J. and A.P.J. Abrahamse, 1970, The correlation coefficient in the general linear model, *European Economic Review* 1, 401–427.
- Press, S.J. and A. Zellner, 1978, Posterior distribution for the multiple correlation coefficient with fixed regressors, *Journal of Econometrics* 8, 307–321.
- Schoenfeld, P., 1969, *Methoden der Oekonometrie*, Band I (Vahlen, Berlin).
- Slater, L.J., 1964, Confluent hypergeometric functions, in: M. Abramovitz and I.A. Stegun, eds., *Handbook of mathematical functions* (Dover, New York) 504–535.
- Theil, H., 1971, *Principles of econometrics* (Wiley, New York).