Department of Economics
UNIVERSITY OF CANTERBURY
CHRISTCHURCH, NEW ZEALAND

ISSN 1171-0705

THE EXACT DISTRIBUTION OF R^2 WHEN THE REGRESSION DISTURBANCES ARE AUTOCORRELATED

Mark L. Carrodus and David E. A. Giles

Discussion Paper

No. 9113

Department of Economics, University of Canterbury
Christchurch, New Zealand

Discussion Paper No. 9113

October 1991

THE EXACT DISTRIBUTION OF R^2 WHEN THE REGRESSION DISTURBANCES ARE AUTOCORRELATED

Mark L. Carrodus and David E. A. Giles

THE EXACT DISTRIBUTION OF R^2 WHEN THE REGRESSION DISTURBANCES ARE AUTOCORRELATED*

Mark L. Carrodus

and

David E.A. Giles

Department of Economics
University of Canterbury

October, 1991

Abstract

This paper provides exact evaluations of the distribution of the usual coefficient of determination when the regression model's errors follow an AR(1) or MA(1) process. This provides insights into the extent to which this measure of goodness of fit is distorted by such model mis-specification.

Address for Correspondence : Professor David E.A. Giles, Department of Economics, University of Canterbury, Christchurch, NEW ZEALAND.

1. Introduction

This paper provides some preliminary results concerning the exact distribution of the coefficient of determination in a regression model which is mis-specified by virtue of the errors being autocorrelated. Both AR(1) and MA(1) disturbances are considered. These results are obtained for a range of data sets, and are compared with their counterparts under serially independent errors.

This type of model mis-specification induces a shift in the distribution of R^2, which in turn alters the probability of observing values of R^2 in any given range. Information of this type is useful to applied researchers, as it assists in the interpretation of a calculated R^2 value when the presence of serial correlation is suspected.

2. Notation and Theory

Consider the model

$$y = X\beta + u \quad ; \quad u \sim N(0,\Omega) \tag{1}$$

where y and u are (n×1); X is (n×k), non-stochastic and of rank k; and β is (k×1). Generally, it is further assumed that $\Omega = \sigma^2 I_n$, so that Ordinary Least Squares (OLS) provides the best linear unbiased estimator of β.

Then, if the model includes an intercept[1], the coefficient of determination can be written unambiguously as

$$R^2 = 1 - \left(\sum_{i=1}^{n} v_i^2 \right) \Big/ \left(\sum_{i=1}^{n} (y_i - \bar{y})^2 \right) , \tag{2}$$

where v_i is the i'th element of the OLS residual vector, $v = y - X(X'X)^{-1}X'y$; y_i is the i'th element of y; and $\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$. More compactly,

$$R^2 = y'(E-M)y/y'Ey, \tag{3}$$

where $M = I_n - X(X'X)^{-1}X'$, $E = I_n - \frac{1}{n}\iota\iota'$, and ι is (n×1) with each element unity.

As Koerts and Abrahamse (1971) show, writing R^2 as a ratio of quadratic forms in the Normal random vector y (as in (3)) facilitates the calculation of its cumulative distribution function (cdf). They calculate the cdf of R^2 for two data sets, assuming $\Omega = \sigma^2 I_n$, and for one data set when Ω corresponds to AR(1) errors.[2]

The c.d.f. of R^2 is

$$F(R^2) = Pr.(R^2 \leq r^2)$$

$$= Pr.[y'(qE-M)y \leq 0] , \qquad (4)$$

where $q = 1 - r^2$. As is well known, after some simple manipulations, we have

$$F(R^2) = Pr.\left[\sum_{j=1}^{n} \lambda_j Z_j^2 \leq 0 \right] , \qquad (5)$$

where the λ_j's are the eigenvalues of $\Omega^{1/2}(qE-M)\Omega^{1/2}$ and the Z_j^2 are independent non-central χ^2 variates, each with one degree of freedom, and with non-centrality parameters given by the squared elements of $P'\Omega^{-1/2}X\beta$, where the columns of P are the eigenvectors corresponding to the λ_j's.

Probabilities of the form (5) can be computed efficiently in various ways. We have used Davies' (1980) algorithm in the SHAZAM package (White et al. (1990)). Having computed $F(R^2)$, numerical differentiation[3] yields the probability density function (pdf) of R^2.

3. Design of the Study

Clearly, the distribution of R^2 depends on X and Ω. We have considered six data sets, n = 20, 60; and AR(1) and MA(1) disturbances. With AR(1) errors $u_t = \rho u_{t-1} + \varepsilon_t$, $|\rho| < 1$, $\varepsilon_t \sim N(0,\sigma_\varepsilon^2)$. With MA(1) errors, $u_t = \varepsilon_t + \theta\varepsilon_{t-1}$, $|\theta| < 1$. The X matrices used are[4]: the annual "spirits" income and price data of Durbin and Watson (1951); the quarterly

Australian Consumers Price Index and its lag; a Normal (30,4) variable and a linear trend; a log-Normal (2.23, 19.58) variable and a linear trend; and the orthogonal regressors $(a_2+a_n)/\sqrt{2}$, $(a_3+a_{n-1})/\sqrt{2}$, where the a_i's are the eigenvectors of the usual "differencing" matrix,[5] A.

Similar data sets have been used in other studies associated with autocorrelation (e.g., Evans (1991)), and a range of characteristics is covered. The last X matrix above is due to Watson (1955) - it produces the least efficient least squares parameter estimates in the class of orthogonal regressor matrices.[6]

Values of $\sigma_\varepsilon^2 = 0.1$, 1.0 and various values of ρ and θ were considered, and the elements of β were controlled to preclude degenerate distributions. The SHAZAM code was checked by replicating the results given by Koerts and Abrahamse (1971, pp.139-140).

4. Results

We concur with previous findings that decreasing σ_ε^2 shifts the cdf (and hence the pdf) of R^2 to the right with serially independent errors. That is, the probability of a low R^2 is decreased. As expected, increasing n concentrates the pdf of R^2. These effects are illustrated in Figures 1 and 2, with $\beta' = (0.001, 0.002, 0.001)$. Both of these results continue to hold with AR(1) or MA(1) errors.

Except for Watsons X matrix, negative AR(1) errors shift the cdf of R^2 increasingly to the left, for any n or σ_ε^2, reflecting a higher probability of underestimating the proportion of total variation explained by the model. Depending on the data, positive AR(1) errors have a mixed effect, contrary to the very limited evidence given by Koerts and Abrahamse (1971, pp.151-152). In particular, the cdf of R^2 does <u>not</u> necessarily shift to the right in this case, though there is a tendency for it to do so.

Contrary to certain econometric folk-lore, positive AR(1) errors do not necessarily introduce a downward bias in the estimation of the error variance.[7] With Watson's X matrix the cdf of R^2 shifts increasingly to the right as the absolute value of ρ increases.

The results with MA(1) errors are even more mixed. With few exceptions, negative autocorrelation of this type shifts the cdf of R^2 to the left. There is no clear pattern regarding such shifts under positive MA(1) autocorrelation. This highlights the importance of having considered a range of data sets. Generally, in this case, the shifts in the cdf of R^2 are less pronounced than in the corresponding positive AR(1) cases, especially with positive autocorrelation. These results are illustrated in Figures 3 and 4, with $\beta' = (0.01, 0.02, 0.02)$, $n = 20$ and $\sigma_\varepsilon^2 = 0.1$.

5. Conclusions

These results have some interesting implications for diligent reporters of R^2. A reasonably large R^2 value is especially encouraging if there is evidence of negative autocorrelation in the errors - such autocorrelation increases the probability of a low R^2. On the other hand, caution is needed in the (likely) presence of positively autocorrelated errors as the likelihood of a high R^2 value is then dependent heavily on the form of the regressor matrix, in an apparently non-systematic way. Work in progress seeks to identify these dependencies, and to determine any possible effects due to multicollinear data.

FIGURE 1

CPI DATA

SERIALLY INDEPENDENT ERRORS

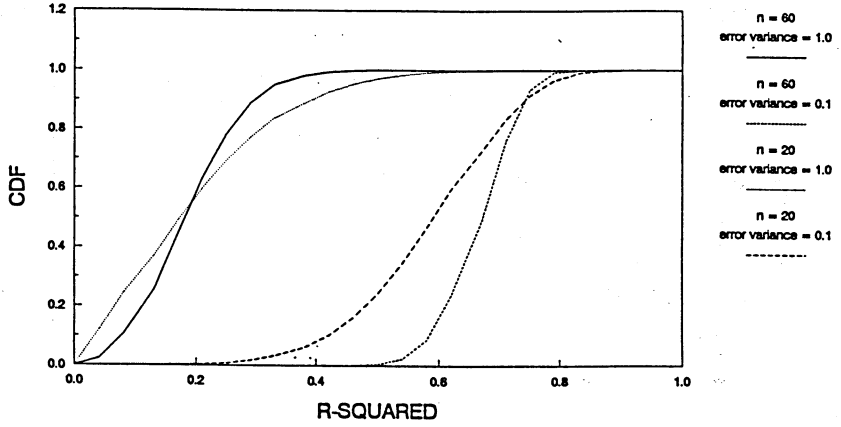


FIGURE 2

CPI DATA

SERIALLY INDEPENDENT ERRORS

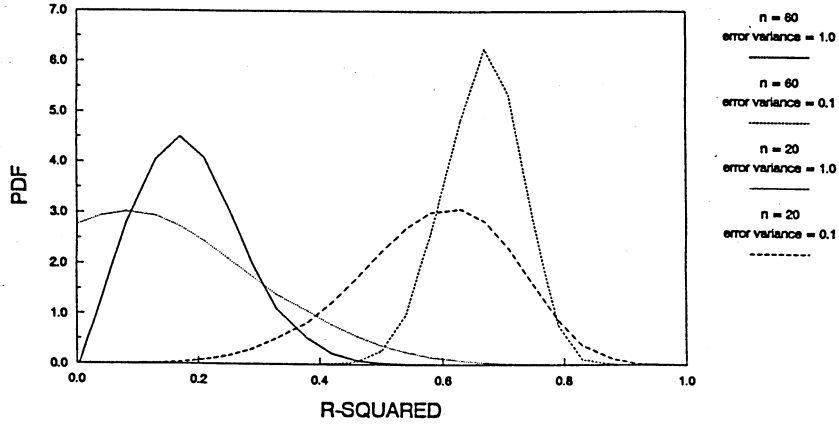


FIGURE 3
LOG-NORMAL & TREND DATA
AR(1) ERRORS

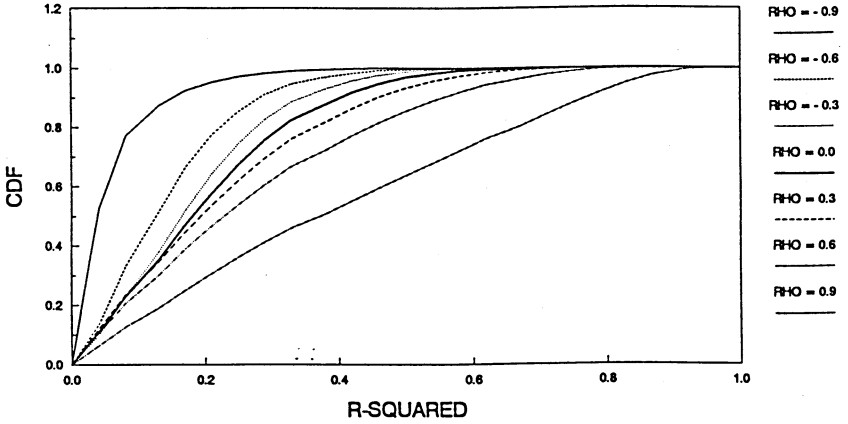
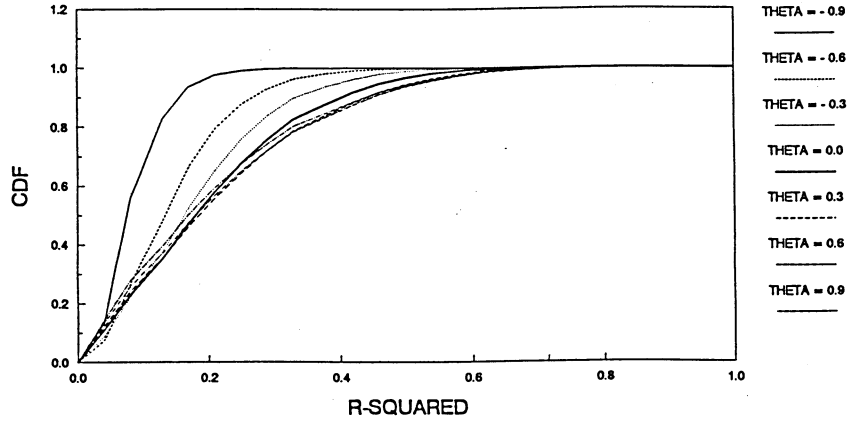


FIGURE 4
LOG-NORMAL & TREND DATA
MA(1) ERRORS



References

Battese, G.E. and W.E. Griffiths, 1980, On R^2-statistics for the general linear model with non-scalar covariance matrix, Australian Economic Papers 19, 343-348.

Cramer, J.S., 1987, Mean and variance of R^2 in small and moderate samples, Journal of Econometrics 35, 253-266.

Davies, R.B., 1980, The distribution of a linear combination of χ^2 random variables: Algorithm AS 155, Applied Statistics 29, 323-333.

Durbin, J. and G.S. Watson, 1951, Testing for serial correlation in least squares regression II, Biometrika 38, 159-178.

Evans, M.A., 1991, Robustness and size of tests of autocorrelation and heteroscedasticity to non-normality, Journal of Econometrics, forthcoming.

Koerts, J. and A.P.J. Abrahamse, 1971, On the theory and application of the general linear model (Rotterdam University Press, Rotterdam).

Nicholls, D.F. and A.R. Pagan, 1977, Specification of the disturbance for efficient estimation - an extended analysis, Econometrica 45, 211-217.

Watson, G.S., 1955, Serial correlation in regression analysis I, Biometrika 42, 327-341.

White, K.J., S.D. Wong, D. Whistler and S.A. Haun, 1990, SHAZAM user's reference manual: Version 6.2 (McGraw-Hill, New York).

Footnotes

* We are grateful to Judith Giles, Murray Scott, John Small and Jason Wong for their helpful comments.

1. If no intercept is included, the value of R^2 depends on whether it is defined as the proportion of "explained" variation, or one minus the proportion of "unexplained" variation in the sample.
2. Cramer (1987) derives expressions for the first two moments of R^2 under certain conditions, and Battese and Griffiths (1980) develop alternative goodness-of-fit measures for the case of a non-scalar error covariance matrix.
3. We have used the method of central differences, with end-point adjustments.
4. Each model also includes an intercept, so $k = 3$ in each case.
5. The matrix A is tri-diagonal, with 2 on the leading diagonal, except for the top left and bottom right elements (which are 1), and -1 on the two leading off-diagonals. The eigenvalues of A are placed in increasing order to number the eigenvectors. The first eigenvector has constant elements.
6. Watson's X matrix is also known to generate extreme situations for the distributions of other statistics (such as the Durbin-Watson statistic) which can be written as ratios of quadratic forms in a Normal vector.
7. Many text book discussions suggest that this is unambiguously the case, but Nicholls and Pagan (1977) provide contrary evidence.