The density functions analysis of R^2 and $\overline{R}^2$ in misspecified linear regression models

Dr. Mahmoud Farouk El-Said

Abstract

In this paper, we analyze the density functions of R^2 and the adjusted R^2 ($\overline{R}^2$) when there are two types of misspecification. The first is exclusion of relevant variables and the other is inclusion of irrelevant variables. It is shown numerically that both R^2 and $\overline{R}^2$ tends to underestimate when there are omitted variables, and both tend to overestimate when there are irrelevant variables.

Introduction:

In applied econometric analysis using regression, the coefficient of determination (say, R^2) and the 'adjusted' R^2 (say, $\overline{R}^2$) are usually reported in the results. Several theoretical analyses have consequently been performed on R^2 and $\overline{R}^2$ For example, Barten [1] suggests a modified version of R^2 to reduce its bias. Press and Zellner [8] discuss the reason why the study of R^2 in the case of fixed regressors is important in econometrics, and perform Bayesian analysis of R^2. Cramer [4] derives the exact first two moments of R^2 and $\overline{R}^2$, and shows that R^2 is seriously biased upward in small samples, and that $\overline{R}^2$ is more unreliable than R^2 in terms of standard deviation, though the bias is relatively small. In practical situations, the model is often misspecified. Although R^2 and $\overline{R}^2$ are usually used as the measures of goodness of fit of the estimated model, studies of their small-sample properties are few when the model is misspecified. Some exceptions are Carrodus and Giles [3], Ohtani [6] and Ohtani and Hasegawa [7]. Carrodus and Giles [3] derive the distribution function of R^2 when the error terms follow an AR(1) or MA(1) process. Ohtani [6] examines the bias and the mean squared error (MSE) of R^2 and an 'improved' R^2 when there are omitted variables. (The 'improved' R^2 is obtained by replacing the ordinary least squares estimator of regression coefficients in the usual R^2 by the so-called Stein rule estimator.) He shows that when the magnitude of specification error is large, both the bias and MSE of the 'improved' R^2 can be larger than those of the usual R^2. Ohtani and Hasegawa [7] examine the bias and MSE of R^2 and $\overline{R}^2$ when proxy variables are used instead of unobservable variables and when the error terms have the normal and the multivariate t distributions. They show that if the unobservable variables are important, $\overline{R}^2$ can be more unreliable than R^2 in small samples in terms of both bias and MSE.

Exclusion relevant variables

Model and estimators:

Suppose that the correct model is

$$y = \ell\beta_0 + X_1\beta_1 + X_2\beta_2 + \varepsilon \qquad \varepsilon \sim N(0,\sigma^2 I_n) \qquad \text{......... } \mathbf{(1)}$$

Where:

$\boldsymbol{y}$: an $n\times 1$ vector of observations, and it represents dependent variable.

ℓ: an $n\times 1$ vector of ones.

$\boldsymbol{X_1}$: an $n \times k_1$ matrix of none stochastic independent variables.

$\boldsymbol{X_2}$: an $n \times k_2$ matrix of none stochastic independent variables.

β_0 : an intercept of regression line.

$\boldsymbol{\beta}_1$: an $k_1 \times 1$ vector of coefficients.

$\boldsymbol{\beta}_2$: an $k_2 \times 1$ vector of coefficients.

$\boldsymbol{\varepsilon}$: an $n \times 1$ vector of normal error terms.

We assume that all independent variables are measures as deviations from their sample mean, $\boldsymbol{X_1}$ and $\boldsymbol{X_2}$ are of full rank.

The model is more compactly written as

$$y = \ell\beta_0 + \mathbf{X}^{\bullet}\beta^{\bullet} + \varepsilon \qquad \textbf{......... (2)}$$

When the researcher omits variables $\boldsymbol{X_2}$ mistakenly, the model is misspecified as

$$y = \ell\beta_0 + X_1\beta_1 + \eta \qquad \textit{where} \qquad \eta = X_2\beta_2 + \varepsilon \qquad \textbf{......... (3)}$$

The ordinary least squares estimators of β_0 and $\boldsymbol{\beta}_1$ based on the misspecified model (3) are

$$b_0 = \bar{y} \qquad \textbf{......... (4)}$$

$$b_1 = S_{11}^{-1}X_1'y \qquad \textit{where} \qquad S_{11} = X_1'X_1 \qquad \textbf{......... (5)}$$

Since the model to be estimated is misspecified as in (3), R^2 is defined as

$$R^2 = \frac{b_1'S_{11}b_1}{b_1'S_{11}b_1 + e_1'e_1} \qquad \textit{where} \qquad e_1 = y - (\ell\bar{y} + X_1b_1) \qquad \textbf{........ (6)}$$

Since the parent coefficient of determination is defined based on the true model given in (2), it is defined as

$$\Phi = \frac{\beta^{\bullet}{}'X^{\bullet}{}'X^{\bullet}\beta^{\bullet}}{\beta^{\bullet}{}'X^{\bullet}{}'X^{\bullet}\beta^{\bullet} + n\sigma^2} \qquad \textbf{......... (7)}$$

Cramer [1987], if we take the probability limit of R^2 when there is no specification error, it reduces to Φ.

The density function:

The adjusted R^2 is defined as

$$\overline{R}^2 = [\tfrac{n-1}{n-k_1-1}]R^2 - [\tfrac{k_1}{n-k_1-1}] \qquad \text{.........}(8)$$

We define the following formally general estimator:

$$R^{\bullet 2} = hR^2 + (1-h) \qquad where\ h \geq 1 \qquad and \quad (1-h) \leq R^{\bullet 2} \leq 1 \ \text{.........}(9)$$

where:

$$R^{\bullet 2} = R^2 \quad when \quad h = 1 \qquad and \qquad R^{\bullet 2} = \overline{R}^2 \qquad when\ h = \tfrac{n-1}{n-k_1-1}$$

Since $R^{\bullet 2}$ can have any value between (1-h) and (1), therefore $\overline{R}^2$ can be negative if

$$R^2 \leq \tfrac{k_1}{n-1}$$

The probability density function of $R^{\bullet 2}$ when there is specification error is defined as the following: **Ohtani [2001]**

$$p(R^{\bullet 2}) = \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} \frac{w_i(\lambda_1)w_j(\lambda_2)}{B(\frac{V_1}{2}+i,\frac{V_2}{2}+j)} h^{(\frac{-(V_1+V_2)}{2}-i-j+1)} (R^{\bullet 2}+h-1)^{(\frac{V_1}{2}+i-1)} (1-R^{\bullet 2})^{(\frac{V_2}{2}+j-1)} \text{.......}(10)$$

Where:

$p(\,)$ is the density function of $R^{\bullet 2}$.

$$w_i(\lambda_1) = \frac{\exp(-\lambda_1/2)(\lambda_1/2)^i}{i!} \qquad where\ \lambda_1 = \frac{\boldsymbol{\beta}^{\bullet\prime}\boldsymbol{X}^{\bullet\prime}\boldsymbol{X}_1\boldsymbol{S}_{11}{}^{-1}\boldsymbol{X}_1{}'\boldsymbol{X}^{\bullet}\boldsymbol{\beta}^{\bullet}}{\sigma^2}$$

$$w_j(\lambda_2) = \frac{\exp(-\lambda_2/2)(\lambda_2/2)^j}{j!} \qquad where \quad \lambda_2 = \frac{\boldsymbol{\beta}^{\bullet\prime}\boldsymbol{X}^{\bullet\prime}\boldsymbol{M}_1\boldsymbol{X}^{\bullet}\boldsymbol{\beta}^{\bullet}}{\sigma^2}$$

$$and\ \boldsymbol{M}_1 = \boldsymbol{I}_n - \frac{\ell\ell'}{n} - \boldsymbol{X}_1\boldsymbol{S}_{11}{}^{-1}\boldsymbol{X}_1{}'$$

$V_1 = k_1 \qquad V_2 = n - k_1 - 1 \qquad B(\frac{V_1}{2}+i, \frac{V_2}{2}+j)$ is beta function.

Numerical results:

- When there is not specification error ($\lambda_2 = 0$), Figure (1) and Figure (2) show that R^2 and $\overline{R}^2$ have upward biases, the upward bias of R^2 is larger than that of $\overline{R}^2$. However, the variance of R^2 is smaller than that of $\overline{R}^2$.

- When there is not specification error ($\lambda_2 = 0$), Figure (3) shows that R^2 has upward biases and $\overline{R}^2$ has downward biases, the upward bias of R^2 is larger than downward bias of $\overline{R}^2$. However, the variance of R^2 is smaller than that of $\overline{R}^2$.

- When there is specification error ($\lambda_2 = 10$), Figure (4) and Figure (5) show that R^2 and $\overline{R}^2$ have downward biases, the downward bias of R^2 is smaller than that of $\overline{R}^2$. However, the variance of R^2 is smaller than that of $\overline{R}^2$.

- When there is specification error ($\lambda_2 = 10$), Figure (6) shows that R^2 and $\overline{R}^2$ have downward large biases, the downward bias of R^2 is larger than that of $\overline{R}^2$. However, the variance of R^2 is smaller than that of $\overline{R}^2$.The variance of R^2 is negative, where the density of R^2 is negative and zero on intervals $[0.15, 0.4]$ and [0.4, 1] respectively.

- Comparing figures (1) and (4), figures (2) and (5) and figures (3) and (6), we see that as specification error increases, the biases of R^2 and $\overline{R}^2$ change the signs from positive to negative, the bias of R^2 becomes smaller than that of $\overline{R}^2$. Since the variance of R^2 is smaller than that of $\overline{R}^2$ irrespective of specification error, therefore the MSE of R^2 is smaller than that of $\overline{R}^2$ as specification error increases.

The all figures, the dashed curve represents the adjusted R^2 ($\overline{R}^2$) and the soled curve represents R^2.

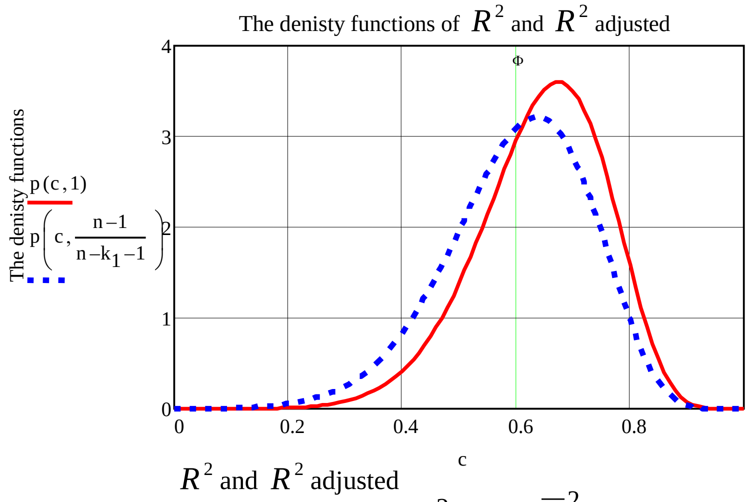

Figure (1): Density functions of R^2 and $\overline{R}^2$ for $n=20$, $k_1=2$, $\Phi=0.6$ and $\lambda_2=0$

$E(R^2)=0.6441$; $Var(R^2)=0.0125$; $E(\overline{R}^2)=0.6022$; $Var(\overline{R}^2)=0.0157$

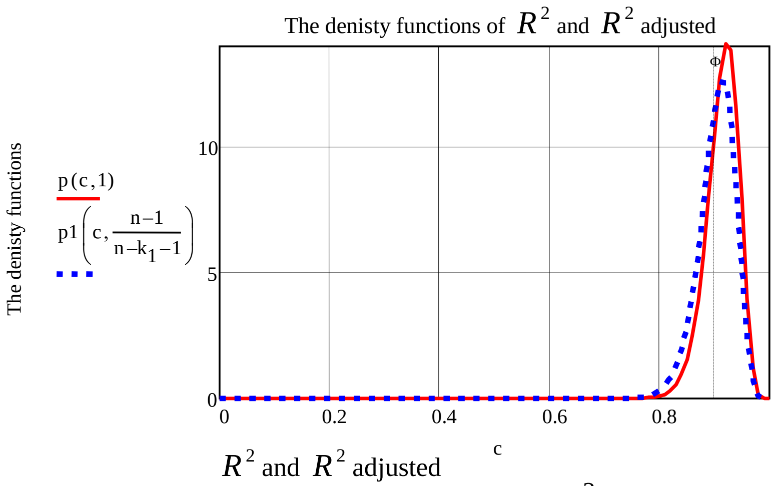

Figure (2): Density functions of R^2 and $\overline{R}^2$ for $n=20$, $k_1=2$, $\Phi=0.9$ and $\lambda_2=0$

$E(R^2)=0.9138$; $Var(R^2)=0.0009$; $E(\overline{R}^2)=0.9036$; $Var(\overline{R}^2)=0.0011$

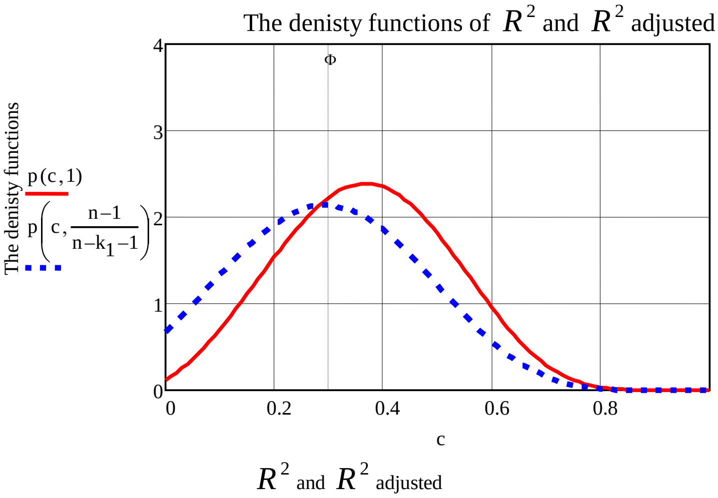

Figure (3): Density functions of R^2 and $\overline{R}^2$ for $n=20$, $k_1=2$, $\Phi=0.3$ and $\lambda_2=0$

$E(R^2)=0.3696$; $Var(R^2)=0.0239$; $E(\overline{R}^2)=0.2972$; $Var(\overline{R}^2)=0.0288$

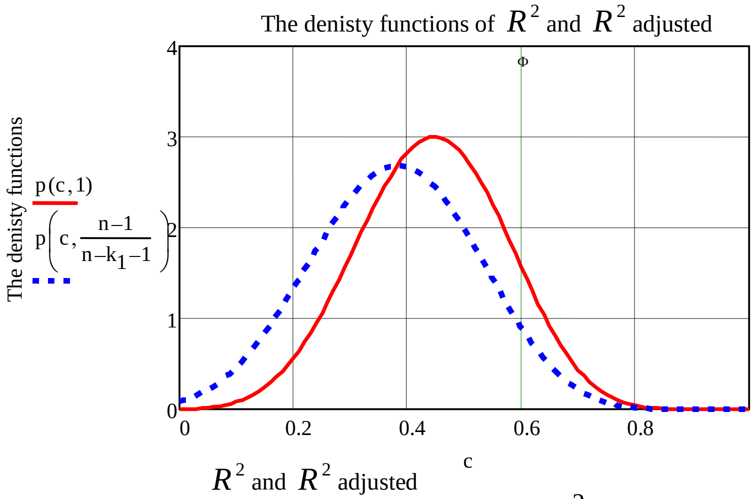

Figure (4): Density functions of R^2 and $\overline{R}^2$ for $n=20$, $k_1=2$, $\Phi=0.6$ and $\lambda_2=10$

$E(R^2)=0.4433$; $Var(R^2)=0.0164$; $E(\overline{R}^2)=0.3779$; $Var(\overline{R}^2)=0.0205$

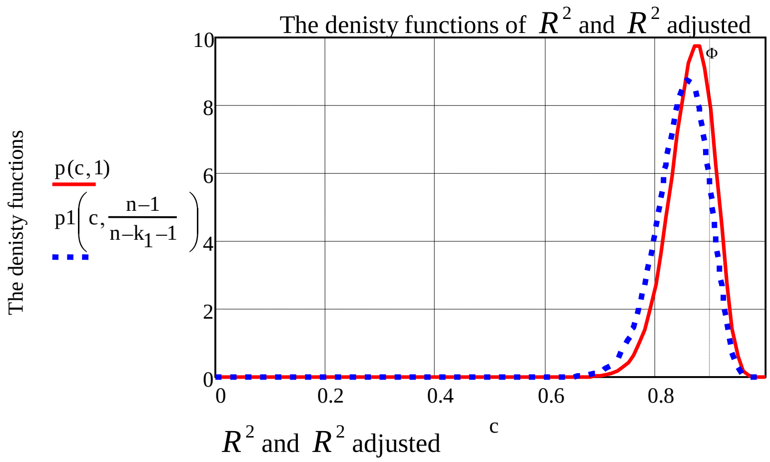

Figure (5): Density functions of R^2 and $\overline{R}^2$ for $n = 20$, $k_1 = 2$, $\Phi = 0.9$ and $\lambda_2 = 10$

$E(R^2) = 0.8636$; $Var(R^2) = 0.0017$; $E(\overline{R}^2) = 0.8475$; $Var(\overline{R}^2) = 0.0022$

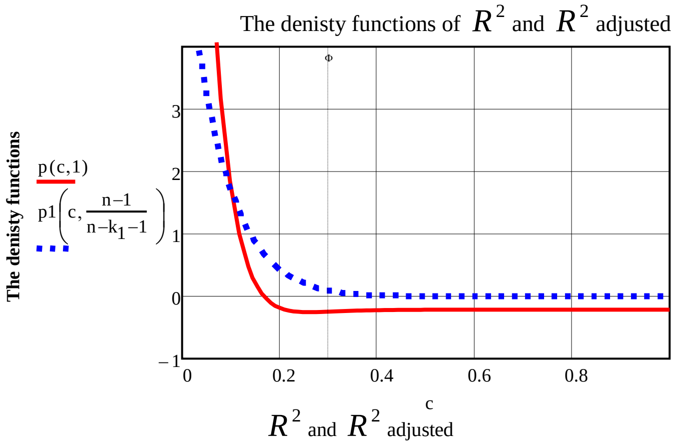

Figure (6): Density functions of R^2 and $\overline{R}^2$ for $n = 20$, $k_1 = 2$, $\Phi = 0.3$ and $\lambda_2 = 10$

$E(R^2) = 0.0249$; $Var(R^2) = -0.0011$; $E(\overline{R}^2) = 0.0343$; $Var(\overline{R}^2) = 0.0036$

Inclusion irrelevant variables:

In a quite parallel way to that above, we can drive the density function of $R^{\bullet 2}$, it is obtained from (10) by replacing V_1 by τ_1 , V_2 by τ_2, λ_1 by μ_1, and λ_2 by 0.

$$p(R^{\bullet 2}) = \sum_{i=0}^{\infty} \frac{w_i(\mu_1)}{B(\frac{\tau_1}{2}+i,\frac{\tau_2}{2})} h^{(\frac{-(\tau_1+\tau_2)}{2}-i+1)} (R^{\bullet 2}+h-1)^{(\frac{\tau_1}{2}+i-1)} (1-R^{\bullet 2})^{(\frac{\tau_2}{2}-1)}$$

Where:

$R^{\bullet 2} = R^2$ *when* $h=1$ *and* $R^{\bullet 2} = \overline{R}^2$ *when* $h = \frac{n-1}{n-k_1-k_2-1}$

$\tau_2 = n - k_1 - k_2 - 1$

$\tau_1 = k_1 + k_2$

k_2 is the number of the irrelevant variables

$$\mu_1 = \frac{\beta_1^{\bullet\prime} \boldsymbol{S}^{\bullet} \beta_1}{\sigma^2}$$

$\mu_1 = \frac{\beta_1^{\bullet\prime} \boldsymbol{S}^{\bullet} \beta_1}{\sigma^2}$ *where* $\beta_1^{\bullet} = (\beta_1', \mathbf{0})'$ *and* $\boldsymbol{S}^{\bullet} = \boldsymbol{X}^{\bullet\prime} \boldsymbol{X}^{\bullet}$

Numerical results:

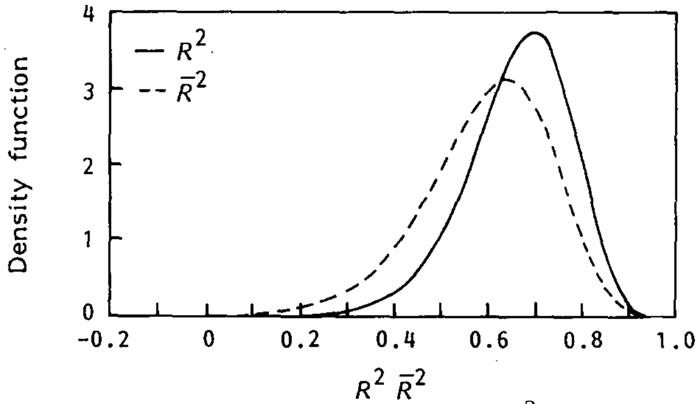

Figure (7): Density functions of R^2 and $\overline{R}^2$ for n = 20, $k_1 = 2$, $\Phi = 0.6$ and $k_2 = 1$

$E(R^2) = 0.665$; $Var(R^2) = 0.0119$; $E(\overline{R}^2) = 0.602$; $Var(\overline{R}^2) = 0.0168$.

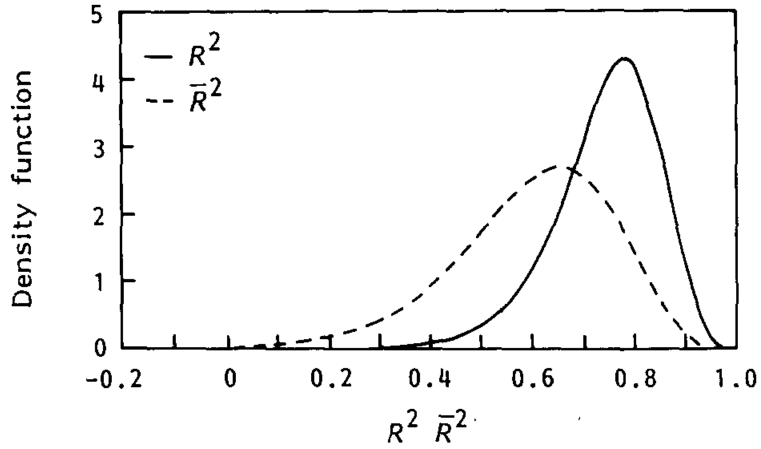

Figure (8): Density functions of R^2 and $\overline{R}^2$ for n = 20, $k_1 = 2$,$\Phi = 0.6$ and $k_2 = 5$

$E(R^2) = 0.749$; $Var(R^2) = 0.0093$; $E(\overline{R}^2) = 0.602$; $Var(\overline{R}^2) = 0.0233$

- Figure (7) shows the density functions of R^2 and $\overline{R}^2$ for n = 20, $k_1 = 2, \Phi = 0.6$ when specification error is small ($k_2 = 1$). We see that both R^2 and $\overline{R}^2$ have upward biases, and the upward bias of R^2 is larger than that of $\overline{R}^2$.

- Figure (8) shows the density functions of R^2 and $\overline{R}^2$ for n = 20, $k_1 = 2$, $\Phi = 0.6$ when specification error is relatively large ($k_2 = 5$). We see that upward bias of R^2 is much larger than that of $\overline{R}^2$, but the variance of R^2 is much smaller than that of $\overline{R}^2$.

Concluding remarks:

In this paper, we have analyzed the density functions of R^2 and $\bar{R}^2$ when there are two types of specification errors for linear regression models.

Our numerical results show the following:

1. When the relevant variables are omitted, and when underestimation is more than overestimation, R^2 is better measure of goodness of fit than $\bar{R}^2$.

2. When irrelevant variables are included, and when underestimation is more than overestimation, R^2 is better measure of goodness of fit than $\bar{R}^2$. When overestimation is more than underestimation, $\bar{R}^2$ is better measure of goodness of fit than R^2.

$$F(c) = P(R_h^2 < c) = \sum_{i=0}^{\infty} w_i(\lambda) I_{c^*}\left(\tfrac{k-1}{2} + i, \tfrac{n-k}{2}\right) \quad .... (7)$$

References:

[1] A.P. Barren, 'Note on the unbiased estimation of the squared multiple correlation coefficient', Statistica Neerlandica, Vol 16, 1962, pp 151-163.

[2] J.O. Berger, Statistical Decision Theory and Bayesian Analysis, 2nd edn, Springer-Verlag, New York, 1985.

[3] M.L. Carrodus and D.E.A. Giles, 'The exact distribution of R^2 when the regression disturbances are auto correlated', Economics Letters, Vo138, 1992, pp 375-380.

[4] J.S. Cramer, 'Mean and variance of R^2 in small and moderate samples', Journal of Econometrics, Vol 35, 1987, pp 253-266.

[5] K. Ohtani, 'The density functions of R^2 and $\overline{R}^2$, and their risk performance under asymmetric loss in misspecified linear regression models, Economic Modelling, Vol 11,1994, pp63-471.

[6] K. Ohtani, 'Small sample properties of R^2 based on the Stein-rule estimator in a misspecified linear regression model', The Economic Studies Quarterly, Vol 44, 1993, pp 263-268.

[7] K. Ohtani and H. Hasegawa, 'On small sample properties of R^2 in a linear regression model with multivariate t errors and proxy variables', Econometric Theory, Vol 9, 1993, pp 504-515.

[8] S.J. Press and A. Zellner, 'Posterior distribution for the multiple correlation coefficient with fixed regressors', Journal of Econometrics, Vol 8, 1978, pp 307-321.