The density functions analysis of $R^2$ and $\bar{R}^2$ 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 $R^2 (\bar{R}^2)$ 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 $\bar{R}^2$ 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' $R^2 (\text{say, } \bar{R}^2)$ are usually reported in the results. Several theoretical analyses have consequently been performed on $R^2$ and $\bar{R}^2$. 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 $\bar{R}^2$, and shows that $R^2$ is seriously biased upward in small samples, and that $\bar{R}^2$ 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 $\bar{R}^2$ 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 $\hat{R}^2$ 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, $\hat{R}^2$ 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

$$y = \ell \hat{\beta}_0 + X_1 \hat{\beta}_1 + X_2 \hat{\beta}_2 + \varepsilon \quad \varepsilon \sim N(0, \sigma^2 I_n) \quad \text{........ (1)}$$

**Where:**

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

$\ell$ : an $n \times 1$ vector of ones.
\(X_1\) : an \(n \times k_1\) matrix of non stochastic independent variables.
\(X_2\) : an \(n \times k_2\) matrix of non stochastic independent variables.
\(\beta_0\) : an intercept of regression line.
\(\beta_1\) : an \(k_1 \times 1\) vector of coefficients.
\(\beta_2\) : an \(k_2 \times 1\) vector of coefficients.
\(\epsilon\) : an \(n \times 1\) vector of normal error terms.

We assume that all independent variables are measures as deviations from their sample mean, \(X_1\) and \(X_2\) are of full rank.

The model is more compactly written as
\[
y = \ell \beta_0 + X^* \beta^* + \epsilon \] ....... (2)

When the researcher omits variables \(X_2\) mistakenly, the model is misspecified as
\[
y = \ell \beta_0 + X_1 \beta_1 + \eta \quad where \quad \eta = X_2 \beta_2 + \epsilon \] ....... (3)

The ordinary least squares estimators of \(\beta_0\) and \(\beta_1\) based on the misspecified model (3) are
\[
b_0 = \bar{y} \] ....... (4)
\[
b_1 = S_{11}^{-1} X_1^* y \quad where \quad S_{11} = X_1^* X_1 \] ....... (5)

Since the model to be estimated is misspecified as in (3), \(R^2\) is defined as
\[
R^2 = \frac{b_1^T S_{11} b_1}{b_1^T S_{11} b_1 + \epsilon_i^T \epsilon_i} \quad where \quad \epsilon_i = y - (\ell \bar{y} + X_1 b_1) \] ....... (6)

Since the parent coefficient of determination is defined based on the true model given in (2), it is defined as
\[
\Phi = \frac{\beta^* X^* X \beta^*}{\beta^* X^* X \beta^* + n \sigma^2} \] ....... (7)

Cramer [1987], if we take the probability limit of \(R^2\) when there is no specification error, it reduces to \(\Phi\).
The density functions analysis of $R^2$ and $\overline{R}^2$ in misspecified linear regression models

**The density function:**
The adjusted $R^2$ is defined as

$$\overline{R}^2 = \left[ \frac{n-1}{n-k_1-1} \right] R^2 - \left[ \frac{k_1}{n-k_1-1} \right]$$  \hspace{1cm} ........(8)

We define the following formally general estimator:

$$R^{*2} = hR^2 + (1-h) \hspace{.5cm} \text{where} \hspace{.5cm} h \geq 1 \hspace{.5cm} \text{and} \hspace{.5cm} (1-h) \leq R^{*2} \leq 1 \hspace{1cm} ........(9)$$

where:

$$R^{*2} = R^2 \hspace{.5cm} \text{when} \hspace{.5cm} h = 1 \hspace{.5cm} \text{and} \hspace{.5cm} R^{*2} = \overline{R}^2 \hspace{.5cm} \text{when} \hspace{.5cm} h = \frac{n-1}{n-k_1-1}$$

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

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

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

$$p(R^{*2}) = \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} \frac{w_i(\lambda_1)w_j(\lambda_2)}{B\left(\frac{p_1}{2} + i, \frac{p_2}{2} + j\right)} h^{\left(-\frac{p_1}{2} - \frac{p_2}{2} - i - j + 1\right)} \left(R^{*2} + h - 1\right)^{\left(\frac{p_1}{2} + i - 1\right)} \left(1 - R^{*2}\right)^{\left(\frac{p_2}{2} + j - 1\right)} \hspace{1cm} ...................(10)$$

Where:

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

$$w_i(\lambda_1) = \frac{\exp\left(-\lambda_1 / 2\right) \lambda_1^i}{i!} \hspace{1cm} \text{where} \hspace{1cm} \lambda_1 = \frac{\beta^* X^* X^{*-1} X_1' X^* \beta^*}{\sigma^2}$$

$$w_j(\lambda_2) = \frac{\exp\left(-\lambda_2 / 2\right) \lambda_2^j}{j!} \hspace{1cm} \text{where} \hspace{1cm} \lambda_2 = \frac{\beta^* X^* M_1 X^* \beta^*}{\sigma^2}$$

and $M_1 = I_n - \frac{X_1' X_1}{n}$

$$V_1 = k_1 \hspace{1cm} V_2 = n - k_1 - 1 \hspace{1cm} B\left(\frac{p_1}{2} + i, \frac{p_2}{2} + j\right) \hspace{1cm} \text{is beta function.}$$
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 \(R^2\).
The density functions analysis of $R^2$ and $\bar{R}^2$ in misspecified linear regression models

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

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

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

$E(R^2) = 0.9138$; $Var(R^2) = 0.0009$; $E(\bar{R}^2) = 0.9036$; $Var(\bar{R}^2) = 0.0011$
The density functions of $R^2$ and $R^2$ adjusted

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

$E(R^2) = 0.3696$; $Var(R^2) = 0.0239$; $E(R^2) = 0.2972$; $Var(R^2) = 0.0288$

The density functions of $R^2$ and $R^2$ adjusted

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

$E(R^2) = 0.4433$; $Var(R^2) = 0.0164$; $E(R^2) = 0.3779$; $Var(R^2) = 0.0205$

191
The density functions analysis of $R^2$ and $\overline{R}^2$ in misspecified linear regression models

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 $R^2$, it is obtained from (10) by replacing $V_1$ by $\tau_1$, $V_2$ by $\tau_2$, $\lambda_1$ by $\mu_1$, and $\lambda_2$ by 0.

\[
p(R^2) = \sum_{i=0}^{\infty} w_i(\mu_1) h^{\frac{-(\tau_1 + \tau_2)}{2}} (R^2 + h - 1)^{\frac{\tau_1}{2} + i - 1} (1 - R^2)^{\frac{\tau_2}{2} - 1}
\]

Where:

$R^2 = R^2$ when $h = 1$ and $R^2 = \bar{R}^2$ 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

$\mu_1 = \frac{\beta_1^* S^* \beta_1}{\sigma^2}$

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

Numerical results:

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

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

193
The density functions analysis of $R^2$ and $\bar{R}^2$ in misspecified linear regression models

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

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

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

- Figure (8) shows the density functions of $R^2$ and $\bar{R}^2$ 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 $\bar{R}^2$, but the variance of $R^2$ is much smaller than that of $\bar{R}^2$. 

194
Concluding remarks:

In this paper, we have analyzed the density functions of $R^2$ and $\overline{R}^2$ 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 $\overline{R}^2$.

2. When irrelevant variables are included, and when underestimation is more than overestimation, $R^2$ is better measure of goodness of fit than $\overline{R}^2$. When overestimation is more than underestimation, $\overline{R}^2$ is better measure of goodness of fit than $R^2$. 
The density functions analysis of $R^2$ and $\bar{R}^2$ in misspecified linear regression models

References:


