The Distribution of the Coefficient of determination in Linear Regression Model: A Review

El Houssainy A. Rady [1], Ahmed Amin El-Sheikh [2] and Soha Othman [3]

[1] Department of Applied Statistics and Econometrics, Institute of Statistical Studies and Research, Cairo University, Egypt
[2] Department of Applied Statistics and Econometrics, Institute of Statistical Studies and Research, Cairo University, Egypt
[3] Department of Applied Statistics and Econometrics, Institute of Statistical Studies and Research, Cairo University, Egypt
[*]Corresponding Author: Elhoussainy@yahoo.com, aham103@yahoo.com, Soha_othman@yahoo.com

Abstract: In this article, we review the different studies about the coefficient of determination in linear regression models and make a highlight about the inferences and the density function of the coefficient of determination which presented under the most common assumption when the error terms obey the normal distributions, and also analyzed the certain effects of departures from normality of the error term

***Keyword*: L**inear Regression Model, Coefficient of Determination R^2 , Exact Distribution, Non-normal Error Term

---

1. Introduction

The linear regression analysis has a prominent role in extracting the statistical information from the data through the determination of relationship between the study and explanatory variables, and the success of linear regression analysis lies on the adequacy of the fitted model in explaining the variations in the data set. A popular tool to determine the suitableness of the fitted model is the coefficient of determination R^2 and its adjusted version $\overline{R^2}$, they are treated as summary measures for the goodness of fit of any linear regression model. The R^2 is based on the proportion of variability of the study variable that can be explained through the knowledge of a given set of explanatory variables.

It should be noted that the value of R^2 does not depend only on the distances between predicted and observed values but also on the variation of the outcome variable. So anything that influences this variation also influences the value of R^2 .

2. Disturbance Normally Distributed

There are many studies concerned with the coefficient of determination which based on the assumption that in many applications of the linear regression model, error terms are assumed normally and independently distributed, each with zero mean and common variance, such as,

Cramer (1987) derived and use easily computable expressions for the mean and variance of R^2 in the standard linear regression model with fixed regressors, and normal disturbances but by employ a slightly non-standard presentation and notation. Assuming that a matrix of regressor variables all have been measured as deviations from their sample means, without touching the definition of y and e.

the familiar measure of goodness of fit is

$$R^2 = \frac{b\acute{X}Xb}{b\acute{X}Xb+\acute{e}e} \quad (1)$$

He derive the density function of R^2 and then its moments by first consider the transformation

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

G is the ratio of two independent non-central chi-square variates, where

$\frac{b\acute{X}Xb}{\sigma^2} \sim \chi^2(\lambda, k-1)$, with ***(k - 1)*** degrees of freedom and non-centrality parameters $\lambda = \frac{\hat{\beta}\acute{X}X\beta}{\sigma^2}$

and

$\frac{\acute{e}e}{\sigma^2} \sim \chi^2(0, m-k)$

Upon introducing this transformation and relabeling the parameters we obtain the density of R^2, with argument r, from the transformation theorem, this yields

$$f(r) = \sum_{j=0}^{\infty} w(j) \frac{1}{\mathrm{B}(u+j, v-u)} r^{u+j-1}(1-r)^{v-u-1} \quad (3)$$

with, $w(j) = \frac{e^{-\frac{1}{2}\lambda}\left(\frac{1}{2}\lambda\right)^j}{j!}$, $u = \frac{1}{2}(k-1)$, $v = \frac{1}{2}(m-1)$

Making use of the properties of the Beta function he obtain the moments of R^2 as

$$E(\mathrm{R}^2) = \sum_{j=0}^{\infty} w(j) \frac{u+j}{\mathrm{v+j}} \quad (4)$$

$$E(\mathrm{R}^2)^2 = \sum_{j=0}^{\infty} w(j) \frac{u+j}{\mathrm{v+j}} \frac{u+j+1}{v+j+1} \quad (5)$$

from the relation that between coefficient of determination R^2 and its adjusted version $\bar{R}^2$ where

$$\bar{R}^2 = (1+h)R^2 - h, \quad h = \frac{k-1}{m-k},$$

he derived the exact formulas for the first two moments of R^2 and $\overline{\mathrm{R}^2}$, and shows that R^2 is seriously biased upward in small samples while $\overline{R^2}$ is more unreliable than R^2 in terms of standard deviation.

Ohtani (1994) derived the exact distribution and density functions of R^2 and adjusted $\overline{R^2}$ assuming that the error term is obey a normal distribution by using the same criteria used in **Cramer (1987)**, and examined their risk performance under asymmetric loss when there are two types of misspecification. One is exclusion of relevant variables and the other is inclusion of irrelevant variables.

It is shown numerically that both R^2 and $\overline{R^2}$ tend to underestimate when there are omitted variables, and both R^2 and $\overline{R^2}$ tend to overestimate when there are irrelevant variables. The risk dominance of R^2 and $\overline{R^2}$depends on whether or not overestimation is more serious than underestimation.

Cheng and et al (2014) used two possible forms of additional information which can be used to obtain consistent estimators of the regression coefficient vector. These two forms are based on the knowledge of the covariance matrix of measurement errors associated with explanatory variables and the knowledge of the reliability matrix of explanatory variables, and Since the form of the conventional is directly related to OLSE of the regression coefficient, he used the two types of available information and obtain an appropriate form of the coefficient of determination which can be used to judge the goodness of a fit in the measurement error models.

Kurz-Kim, Loretan (2014) examined the asymptotic properties of the coefficient of determination, R^2, in models with α-stable random variables. If the regressor and error term share the same index of stability $\alpha < 2$, they show that the R^2 statistic does not converge to a constant but has a nondegenerate distribution on the entire [0, 1] interval. They provide closed-form expressions for the cumulative distribution function and probability density function of this limit random variable, and the density function is unbounded at 0 and 1.

3. Disturbance Non-Normally Distributed

Since some economic data may be distributed through different distributions other than the normal distribution, some studies examined investigated the small sample properties of R^2 and $\overline{R^2}$ when the error term is distributed non-normal distribution, but they are still based on the asymptotic distribution such as:

Arnold Zellner (1993) analyzed the linear multiple regression model assuming the error vector has a multivariate Student-t distribution with zero location vector and scalar dispersion matrix.

The basic criteria that he used are to rewrite the multivariate Student-t pdf as a member of the class of distributions $$p(u|\cdot) = \int_0^\infty p_N(u|\tau)p(\tau|\cdot)d\tau \quad (6)$$

where

$p_N(u|\tau) = (2\pi\tau^2)^{-n/2}exp(-úu/2\tau^2), \quad -\infty < u_i < \infty, \quad i = 1, 2, \cdots, n$

$p(\tau|\cdot)$ with $0 < \tau < \infty$ is a proper pdf for τ

Assuming that $p(\tau|\cdot)$ is chosen to be the inverted gamma pdf

$$p_{IG}(\tau|v_0, \sigma) = [2/\Gamma(v_0/2)][v_0\sigma^2/2]^{v_0/2\tau-(v_0+1)} \cdot \exp(-v_0\sigma^2/2\tau^2)$$

with $0 < v_0, \sigma, \tau < \infty$

Thus, it is possible to regard the error vector u as being randomly drawn from a multivariate normal distribution with a random standard deviation generated from inverted gamma pdf, and then by integration in (6) produces a marginal multivariate Student-t pdf for u

It is found that the usual least squares coefficient estimate is the maximum likelihood. Inferences based on usual t- and F-statistics are shown valid for a range of error distributional assumptions including the multivariate-t assumption. Thus, while ML estimates of the parameters exist for any given $\boldsymbol{v_0}$, ML estimates of the parameters and $\boldsymbol{v_0}$ do not exist. Thus, it is necessary to assign a value to $\boldsymbol{v_0}$ that reflects an investigator's knowledge of the distributional properties of the regression error terms.

In the Bayesian analysis of the model with a diffuse prior pdf for the regression coefficients and multivariate Student-t error terms, it was found that the joint posterior distribution for the regression coefficients is in precisely the same multivariate Student-t form as arises from the usual normal model.

Ohtani and Hasegawa (1993) examined the bias and mean squared error (MSE) performances when the proxy variables are used instead of unobservable regressors and the error terms obey a multivariate t distribution. Using the criteria using in Arnold Zellner article (1993), the results show that if the unobservable variable is an important variable, the adjusted coefficient of determination can be more unreliable in small samples than the unadjusted coefficient of determination from both viewpoints of the bias and the MSE.

Srivastava and Ullah (1995) examined the sampling properties of R^2 and $\overline{R^2}$ under a general non-normal error distribution, their analysis is based on the large sample asymptotic expansions.

Ohtani and Tanizaki (2004) consider a linear regression model when error terms obey a multivariate t distribution, and derive the exact formulas for the density function, distribution function and $m-th$ moment, and shown that the upward bias of R^2 get serious and the standard error of R^2 gets large as the degrees of freedom of the multivariate t error distribution get small.

REFERENCES

1. A.K. Srivastava, V.K. Srivastava, A. Ullah , "The coefficient of determination and its adjusted version in linear regression models", *Econometric Reviews*, 14, **(1995)**, pp. 229–240.
2. C. L. Cheng, G.G. Garg, "Coefficient of determination for multiple measurement error models", *Journal of Multivariate Analysis*, 126, **(2014)**, pp. 173-152.
3. J.S.Cramer, "Mean and variance of R^2 in small and moderate samples", *J. Econometrics*, 35, **(1987)** , pp. 253–266.
4. K. Ohtani, H. Hasegawa , "On small scale properties of R^2 in a linear regression model with multivariate t errors and proxy variables", *Econom. Theory*, 9, **(1993)** , pp. 504–515.
5. K. Ohtani, "The density functions of R^2 and, $\bar{R}^2$ and their risk performance under asymmetric loss in misspecified linear regression models", *Econom. Model*, 11, **(1994)** , pp. 463–471.
6. K. Ohtani and H. Tanizaki , "Exact Distributions of R^2 and Adjusted R^2 in a Linear Regression Model with Multivariate t Error Terms", *Journal of the Japan Statistical Society*, 1,**(2004)** , pp. 101-109.
7. Kurz-Kim, Mico Loretan, " On the properties of the coefficient of determination in regression models with infinite variance variables", J*ournal of Econometrics*, 1, **(2014)** , pp. 15-24.
8. A. Zellner, Bayesian and non-Bayesian analysis of the regression model with multivariateStudent-t error terms. Journal of the American Statistical Association, 71, **(1976)** , pp. 400-405.