Journal of Multivariate Analysis 126 (2014) 137–152

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

Coefficient of determination for multiple measurement error models[☆]

C.-L. Cheng[a], Shalabh[b,*], G. Garg[c]

[a] Institute of Statistical Science, Academia Sinica, Taipei, Taiwan, ROC
[b] Department of Mathematics and Statistics, Indian Institute of Technology Kanpur, Kanpur - 208 016, India
[c] Decision Sciences Area, Indian Institute of Management Lucknow, Lucknow - 226 013, India

ARTICLE INFO

Article history:
Received 27 December 2012
Available online 5 February 2014

AMS subject classifications:
62J05
62H12

Keywords:
Measurement error
Linear regression
Coefficient of determination (R^2)
Ultrastructural model
Non-normal distribution

ABSTRACT

The coefficient of determination (R^2) is used for judging the goodness of fit in a linear regression model. It is the square of the multiple correlation coefficient between the study and explanatory variables based on the sample values. It gives valid results only when the observations are correctly observed without any measurement error. The conventional R^2 provides invalid results in the presence of measurement errors in the data because the sample R^2 becomes an inconsistent estimator of its population counterpart which is the square of the population multiple correlation coefficient between the study and explanatory variables. The goodness of fit statistics based on the variants of R^2 for multiple measurement error models have been proposed in this paper. These variants are based on the utilization of the two forms of additional information from outside the sample. The two forms are the known covariance matrix of measurement errors associated with the explanatory variables and the known reliability matrix associated with the explanatory variables. The asymptotic properties of the conventional R^2 and the proposed variants of R^2 like goodness of fit statistics have been studied analytically and numerically.

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. An adequate linear regression model provides valid statistical inferences on various applications including the forecasting. 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 adequacy of the fitted model is the coefficient of determination and the adjusted version. The coefficient of determination is popularly known as R^2 and its adjusted version is called as adjusted 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 is the square of the multiple correlation coefficient between the study variable and all the explanatory variables present in the linear regression model. The R^2 and its adjusted version are also used for the model selection. For example, if there are several fitted models available from the same data set, then

[☆] The authors are grateful to the referees for their comments to improve the exposition of the paper.
[*] Corresponding author.
E-mail addresses: clcheng@stat.sinica.edu.tw (C.-L. Cheng), shalab@iitk.ac.in, shalabh1@yahoo.com (Shalabh), ggarg@iiml.ac.in, ggarg31@gmail.com (G. Garg).

http://dx.doi.org/10.1016/j.jmva.2014.01.006
0047-259X/© 2014 Elsevier Inc. All rights reserved.

a model with the least lack of fit is preferred and can be determined based on the values of the coefficient of determination or its adjusted version. Although R^2 and its adjusted versions have certain limitations, see [12], but in spite of them, they remain a popular choice among practitioners.

The research work in obtaining the different suitable forms of the coefficient of determination for various situations has been addressed in the literature by several researchers. Eshima and Tabata [6,7] proposed the coefficient of determination in entropy form for generalized linear models. Renaud and Victoria-Feser [28] presented a robust coefficient of determination in regression. Tjur [42] proposed a coefficient of determination for the logistic regression model, see also [14,18]. Huang and Chen [16] addressed the issue of the coefficient of determination in the local polynomial model. Hössjer [15] discussed the role of the coefficient of determination in the mixed regression model. Linde and Tutz [44] considered the coefficient of determination in the case of association in a regression framework. Srivastava and Shobhit [39] proposed a family of coefficients of determination in the linear regression model. Marchand [21] discussed the point estimation of the coefficient of determination, see also [22]. Lipsitz et al. [19] discussed the partial correlation coefficient and the coefficient of determination for the multivariate normal repeated measures data. Tanaka and Huba [41] presented a general coefficient of determination for the covariance structure models under arbitrary generalized least squares estimation. Nagelkerke [24] presented a generalization of the coefficient of determination. McKean and Sievers [23] obtained a new coefficient of determination for the least absolute deviation analysis. Knight [17] and Hilliard and Lloyd [13] discussed the role of the coefficient of determination in the simultaneous equation models. Ohtani [25] derived the density of R^2 and its adjusted version. He also analyzed their risk performance under an asymmetric loss function in the misspecified linear regression model.

One of the fundamental assumptions in using the coefficient of determination in the linear regression analysis is that all the observations on the study and explanatory variables are correctly observed. Many times in practical situations, the variables are not correctly observable and the measurement errors creep into the data. If the magnitude of measurement errors is negligible, then it may not pose any big challenge to the derived statistical inferences. On the other hand, when the magnitude of measurement errors is large, then it disturbs the optimal properties of the estimators. A serious consequence of measurement errors in linear regression analysis is that the ordinary least squares estimator (OLSE) of the regression coefficients becomes biased and inconsistent. Note that the same OLSE is the best linear unbiased estimator of the regression coefficients in the absence of measurement errors in the data. The coefficient of determination (R^2) is a function of OLSE. So consequently, the presence of the measurement error disturbs the properties of R^2. The value of R^2 obtained by ignoring the measurement errors becomes misleading and may provide incorrect statistical inferences. So we are faced with the question of how to judge the goodness of fit in the linear regression model when the observations are contaminated with measurement errors. Such an issue has never been addressed in the literature, to the best of our knowledge.

It may also be noted that the expression of conventional R^2 in the multiple linear regression model is based on the analysis of variance. It is defined as the ratio of the sum of squares due to regression and the total sum of squares. Unfortunately, such analysis of variance in the setup of measurement error models is not possible. This is due to the nonexistence of the moments of the estimators and the complicated structure of moments, if they exist in some cases, see [2,1]. So the only option left is possibly to look at the structure of R^2 and adjust it in the framework of the measurement error model so as to reflect the goodness of fit. We have attempted in this direction.

In order to obtain the consistent estimators of regression coefficients in the presence of measurement errors in the data, the OLSE is adjusted for its inconsistency. Such an adjustment is done by using the additional information from outside the sample. Various forms of additional information can be used to obtain the consistent estimators, see [3,8] etc. for more details. In the context of the multiple measurement error model, there are 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, see, e.g. [3,29,9,10,31–33] etc. Since the form of the conventional R^2 is directly related to OLSE of the regression coefficient in the no-measurement error linear regression model, an idea to obtain statistics for judging the goodness of fit in the measurement error model can be based on the form of conventional R^2. Our objective in this paper is to use both 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.

The plan of the paper is as follows. The multivariate ultrastructural model and the various statistical assumptions are described in Section 2. In Section 3, we demonstrate the inconsistency of the coefficient of determination under the ultrastructural form of the measurement error model. We propose two goodness of fit statistics based on R^2-like expressions. These statistics are consistent for the population counterpart of R^2 which is the square of the population multiple correlation coefficient. The asymptotic distributions of the proposed R^2 like goodness of fit statistics are derived under the specification of the ultrastructural measurement error model in Section 5. In order to study the small sample properties of the proposed goodness of fit statistics, Monte Carlo simulation experiments are conducted. The findings of the simulation study are presented in Section 6 followed by some concluding remarks in Section 7. Lastly, the proof of the results is given in the Appendix.

2. The model

We consider the following exact relationship between the $(n \times 1)$ vector of values of study variable η and the $(n \times p)$ matrix Ξ of n values on each of the p explanatory variables:

$$\eta = \alpha \mathbf{1}_n + \Xi \beta, \tag{2.1}$$