Journal of Multivariate Analysis 145 (2016) 101–116

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

Goodness of fit in restricted 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 22 April 2014
Available online 18 December 2015

AMS (2010) subject classifications:
62J05
62H12

Keywords:
Coefficient of determination
Goodness of fit
Least squares estimator
Linear equality constraint
Measurement error
Restricted regression
Ultrastructural model

ABSTRACT

The restricted measurement error model is employed when certain study variables are not observable by direct measurement and if some information about the unknown regression coefficients is available a priori. In this study, we present a method for checking the goodness of fit in the restricted measurement error model. We obtain the goodness-of-fit statistics based on the concept of coefficient of determination and their asymptotic distributions are derived. The results of simulations are also presented to demonstrate the finite sample behaviour of the estimators.

1. Introduction

A basic assumption of linear regression analysis is that all observations of the study variable and the explanatory variables are correctly observed. Under this condition and the other usual assumptions for a linear regression model, the ordinary least squares (OLS) estimator provides the best linear unbiased estimate of the regression coefficients (see [32]). However, in many practical situations, the observations of the variables are subject to measurement errors, and thus the OLS estimators lose their optimality properties so they become biased and inconsistent when estimating the regression coefficients. This lack of consistency can be addressed when some additional information is available from external sources (see [4,10] for more details). In this study, we distinguish between classical linear regression (CLR) models (without measurement errors) and linear regression models with measurement errors (ME models).

We consider two different types of model constraint: the first type enters the model through external a priori information, which is assumed to be available, e.g., from similar studies conducted previously or through the experience gained by the experimenter; and the second type of constraint enters in the form of analytic (in)equalities on the parameters involved. The former type is called "a-priori information" and the latter are known simply as "constraints". A specific example for a constraint is given by the Cobb–Douglas function in economics, which describes "constant returns to scale" in special cases. Similar constraints can be found in finance research and many other application areas.

☆ The authors are grateful to the associate editor and referees for their useful comments, which improved this paper. The first author's research was partly supported by the Ministry of Science and Technology of Taiwan, Republic of China.

* 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.2015.12.005

There are various types of constraints, such as linear or non-linear (in)equalities, which can be stochastic or non-stochastic (see [52,32] for more details). Restricted least squares (RLS) estimators can be used for CLR models with equality constraints, which are known to be unbiased and more efficient than OLS estimators. However, these properties of RLS estimators do not hold in ME models, where the RLS estimators become unbiased and inconsistent. To address these shortcomings, consistent and more efficient estimators have been proposed for use in ME models by assuming the availability of external a-priori information (e.g., see [39–42,18,44,20]). These estimators tend to be asymptotically normal, but the form of their asymptotic covariances is usually quite complex (e.g., see [43]), which makes it challenging to compare estimators constructed under different types of external a-priori information.

In the present study, we consider the problem of assessing the goodness of fit. The traditional measure of goodness of fit in CLR models is the coefficient of determination, R^2, but this classical quantity is not appropriate for ME models because the underlying estimators are inconsistent in these models. It should be noted that the derivation of the classical R^2 is based on the analysis of variance techniques. The finite sample distributions of consistent estimators in ME models are very difficult to analyse, and thus the derivation of analogues of R^2 in ME models requires different approaches.

The search for suitable goodness-of-fit statistics under various models has been a popular topic in statistics. The coefficient of determination R^2 for generalized linear models in entropy form was discussed by Eshima and Tabata [8,9]; for logistic regression by Tjur [51], Hong et al. [15], and Liao and McGee [21]; for local polynomial models by Huang and Chen [17]; and for mixed regression models by Hössjer [16]. In linear regression models, R^2 was discussed by Linde and Tutz [54], Srivastava and Shobhit [48], and Marchand [24,25]. The generalization of R^2 was presented by Nagelkerke [27], whereas Tanaka and Huba [50] presented R^2 for covariance structure models under arbitrary generalized least squares estimation. R^2 for the least deviation analysis was discussed by McKean and Sievers [26] and R^2 for simultaneous equation models was discussed by Knight [19] and Hilliard and Lloyd [14] (also see [22] for the use of the partial correlation coefficient and the coefficient of determination in repeated data). A robust R^2 in regression was studied by Renaud and Victoria-Feser [33]. The properties of R^2 were studied by Cramer [5]. The relative performance of R^2 for an absolute error loss function was studied by Ohtani and Giles [29], whereas the properties of R^2 under the t-distribution of errors for mis-specified models were studied by Ohtani and Hasegawa [30]. The distribution of R^2 and its adjusted version were studied by Ohtani [28] for asymmetric loss in mis-specified linear models. The small disturbance asymptotic theory was utilized by Ullah and Srivastava [53] to approximate the moments of R^2. Later, Srivastava et al. [49] also employed the large sample asymptotic approximation theory to find the efficiency properties of R^2. The closeness of properties of R^2 under small disturbance approximations and asymptotic approximations were discussed by Smith [45]. Cheng et al. [2] proposed goodness-of-fit statistics based on the idea of R^2 for ME models.

In this study, we present some goodness-of-fit statistics in ME models using a-priori information about unknown regression coefficients that can be expressed as equality constraints. These goodness-of-fit statistics are based on the classical concept of R^2. In this study, we consider the ultrastructural form of the multiple ME model, where the measurement errors are not necessarily normally distributed. The asymptotic properties of the proposed goodness-of-fit statistics are derived and analysed, while we also study the finite sample performance of the proposed statistics based on Monte Carlo simulations. The remainder of this paper is organized as follows. First, we present the multiple ultrastructural model and equality constraints in Section 2. The estimators under measurement errors and equality constraints are described in Section 3. The goodness-of-fit statistics are developed in Section 4 and their asymptotic properties are reported in Section 5. The results obtained from Monte Carlo simulations are discussed in Section 6. Finally, we give some concluding remarks in Section 7.

2. Model and equality constraints

Let η be the $(n \times 1)$ vector of true but unobservable observations of the study variable and Ξ be the $(n \times p)$ matrix of n true but unobservable observations of each of the p explanatory variables. Let η and Ξ have the following relationship:

$$\eta = \alpha e_n + \Xi\beta, \tag{2.1}$$

where α is the intercept term, e_n is an $(n \times 1)$ vector with all unity elements, and β is the $(p \times 1)$ vector of regression coefficients. Let ξ_{ij} be the i-th observation of the j-th explanatory variable such that $\Xi = ((\xi_{ij}))$, $i = 1, \ldots, n; j = 1, \ldots, p$. The presence of an intercept term in the model is required for the goodness-of-fit statistics proposed later in this study. The observed values of the study and explanatory variables are contaminated with measurement errors, so the observed study variable y is written as

$$y = \eta + \varepsilon, \tag{2.2}$$

and the observed explanatory variable X as

$$X = \Xi + \Delta. \tag{2.3}$$

Without loss of generality, the usual disturbance in the linear regression model is assumed to be incorporated in ε, where $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^\top$ is an $(n \times 1)$ vector of measurement errors involved in the $(n \times 1)$ vector of observations of the study variable y and $\Delta = ((\delta_{ij})) = (\delta_1, \ldots, \delta_n)^\top$ is an $(n \times p)$ matrix of measurement errors involved in the observed explanatory