한국통계학회논문집 제3권 2호
1996년 8월 pp. 21-28

On Alternative Collinearity Diagnostics in Linear MEM[1)]

Myung-Sang Moon[2)]

Abstract

Collinearities contained in MEM cause the same problems as they do in traditional regression model, so the detection of collinearities is a crucial topic in MEM. One diagnostic was introduced by Carrillo-Gamboa and Gunst, but their method did not work in some cases. Two alternative collinearity diagnostics that provide reasonable measure of collinearities are proposed. Simulation study is performed to compare the small-sample properties of the proposed collinearity diagnostics.

1. Introduction

In traditional multiple linear regression models in which predictor variables are assumed to be nonstochastic and error-free, ordinary least squares(OLS) estimation method is frequently used to estimate regression coefficients. Despite possessing very desirable properties under the usual conditions of the model, OLS estimators can have extremely large variances if a severe collinearity problem is included in the model. Hence, identifying the existence of a severe collinearity problem has been a crucial topic in traditional regression area, and much research(Belsley, Kuh and Welsch 1980; Gunst 1983; Silvey 1969) on developing collinearity diagnostics has been conducted.

In many situations, some or all of predictor variables in a regression model are measured with error and that model is referred to as measurement error model(MEM). When predictor variables are subject to error, usual OLS estimator used with traditional regression models is inconsistent, and alternative estimation methods have been developed. Collinearities can occur with MEM fit by any estimation method, and they have the ill effects on the regression coefficient estimators as they do on OLS estimator in the traditional regression. So, detecting reasonable collinearity diagnostics is one of important issues in MEM. Nevertheless, a few diagnostic procedure for collinearities is currently available although there is a voluminous literature on parameter estimation. Jagpal (1982) developed a ridge estimator for the treatment of collinearity problems in MEM. Carrillo-Gamboa and Gunst (1992) mentioned extensively on

1) This Research was supported by the 1995-1996 Yonsei Maeji Research Fund.
2) Assistant Professor, Department of Statistics, Yonsei University, Wonju-City, Kangwon-Do, 222-701, Korea.

the problems of collinearities in MEM including definition, inducing and masking of collinearities and detection of collinearities. They defined the collinearities in terms of second-order moment matrix of unobservable error-free predictors(defined in section 2) for linear MEM and proposed one diagnostic based on the smallest eigenvalue of estimated second-order moment matrix of error-free predictors(defined in section 2). But their diagnostic did not always provide reasonable measure of collinearities, since it took negative values in some cases as was shown in their simulation results. As a possible alternative, they suggested a use of modified second-order moment matrix such as proposed by Amemiya (1985). However, Amemiya's method does not work either since his suggested modified second-order moment matrix is always positive semidefinite and its smallest eigenvalue is zero.

The purpose of this work is to detect reasonable alternative collinearity diagnostics that eliminate the problems of Carrillo-Gamboa and Gunst's diagnostic. They are supposed to provide absolutely positive smallest eigenvalue of estimated second-order moment matrix of error-free predictors. In section 2, some basic notation needed in this work is introduced, and Carrillo-Gamboa & Gunst's results are briefly summarized. Two alternative collinearity diagnostics and their properties are presented in section 3. Simulation scheme comparing the small-sample properties of suggested diagnostics is described in section 4. Final section is devoted to concluding remarks that describe the simulation results.

2. Collinearities in MEM

In MEM, true response and true predictor variables are unobservable and they are contaminated as follows due to measurement errors:

$$y_i = \phi_i + v_i, \quad x_{ij} = \pi_{ij} + u_{ij}, \quad j = 1, 2, 3, \cdots, k, \quad i = 1, 2, 3, \cdots, n.$$

Let $\boldsymbol{z}_i = (y_i, \boldsymbol{x}_i^t)^t$ be the i^{th} vector of observed response and predictor variables; i.e., $\boldsymbol{z}_i = \boldsymbol{\xi}_i + \boldsymbol{w}_i$ with $\boldsymbol{\xi}_i = (\phi_i, \boldsymbol{\pi}_i^t)^t$ denoting the i^{th} vector of error-free variates, and $\boldsymbol{w}_i = (v_i, \boldsymbol{u}_i^t)^t$ being the i^{th} vector of measurement errors. Bold-faced letters denote vectors or matrices and all vectors are column ones in this work. Assume that measurement errors $\boldsymbol{w}_i$ are iid $MVN(\boldsymbol{0}, \boldsymbol{\Sigma}_{ww})$, where $\boldsymbol{\Sigma}_{ww}$ is positive semidefinite.

In a linear MEM, an unobservable response is a linear function of unobservable predictors. That is,

$$\phi_i = \boldsymbol{\pi}_i^t \boldsymbol{\beta}, \tag{2.1}$$

where β is a $(k+1)$ x 1 vector of regression coefficients. The observable model is $y_i = x_i^t \beta + e_i$, where $e_i = v_i - u_i^t \beta$. The model in (2.1) is called a no-equation-error model(Fuller 1987). An equation-error representation of the model is $\psi_i = \pi_i^t \beta + q_i$, where q_i denotes the error in the equation. If π_i's in model (2.1) represent a sequence of constant vectors, then the model is defined as a functional MEM. In a structural MEM, error-free predictor variables are stochastic. Assumptions needed for a functional model include the existence of following first (μ_x) and second-order (Γ_{xx}) moment matrix of error-free predictors:

$$\mu_x = \lim_{n\to\infty} n^{-1} \sum_{i=1}^{n} \pi_i \quad \text{and} \quad \Gamma_{xx} = \lim_{n\to\infty} n^{-1} \sum_{i=1}^{n} \pi_i \pi_i^t = \lim_{n\to\infty} \Gamma_{nn}, \tag{2.2}$$

where Γ_{xx} is positive definite. Note that these limits also hold for a structural model with the assumption that $\Gamma_{xx} = \Sigma_{xx} + \mu_x \mu_x^t$ is positive definite.

MEM considered in this work is no-equation-error linear functional model with $k+1$ predictors including intercept term. When the error covariance matrix is known upto a multiple, that is $\Sigma_{ww} = \sigma^2 \mathbf{T}_{ww}$ with $\mathbf{T}_{ww}$ known, the estimated second-order moment matrix of error-free predictors is always positive definite. Therefore, collinearity diagnostic suggested by Carrillo-Gamboa and Gunst has no problem in detecting collinearities as was mentioned in their paper. Since we are supposed to deal with the case where their collinearity diagnostic does not work, the error covariance matrix Σ_{ww} is assumed to be completely known or completely unknown in this work.

Let $M_{zz} = n^{-1} \sum_{i=1}^{n} z_i z_i^t = \begin{pmatrix} M_{yy} & M_{xy}^t \\ M_{xy} & M_{xx} \end{pmatrix}$. Similarly, partition Σ_{ww} to conform with w so that $\Sigma_{ww} = \begin{pmatrix} \sigma_{vv} & \Sigma_{uv}^t \\ \Sigma_{uv} & \Sigma_{uu} \end{pmatrix}$. Assuming no-equation-error model and completely known Σ_{ww}, a regression coefficients estimator is

$$\hat{\beta} = (M_{xx} - \hat{\lambda}\ \Sigma_{uu})^{-1}(M_{xy} - \hat{\lambda}\ \Sigma_{uv}), \tag{2.3}$$

where $\hat{\lambda}$ is the smallest root of $|M_{zz} - \lambda \Sigma_{ww}| = 0$. Under the assumed model of this work, $n^{1/2}(\hat{\beta} - \beta)$ is asymptotically normal with zero mean vector and covariance matrix given by

$$\Omega_{\beta\beta} = \Gamma_{xx}^{-1} \{ (\Gamma_{xx} + \Sigma_{uu}) \sigma_{ee} - \Sigma_{ue} \Sigma_{ue}^{t} \} \Gamma_{xx}^{-1}. \tag{2.4}$$

The asymptotic variance formula (2.4) involves the inverse of second-order moment matrix of error-free predictors (Γ_{xx}), so Γ_{xx} in MEM plays the role of $X^t X$ matrix in traditional regression model. For this reason, collinearities are a property of the second-order moment matrix of error-free predictors for linear MEM. However, we should have reasonable estimator of Γ_{xx} to assess the degree of collinearities included in the model, since Γ_{xx} involves the unobservable error-free predictor variables. Carrillo-Gamboa and Gunst suggested to use a method-of-moments estimator

$$\tilde{\Gamma}_{xx} = M_{xx} - \Sigma_{uu} \tag{2.5}$$

for no-equation-error model with completely known Σ_{ww} or with completely unknown but unbiasedly estimated Σ_{ww}. Unfortunately, $\tilde{\Gamma}_{xx}$ is not always positive definite and sometimes its smallest eigenvalue is negative. Hence, their collinearity diagnostic failed to provide a reasonable measure of collinearities in some cases. In the next section, two alternative collinearity diagnostics whose smallest eigenvalues are always positive, are proposed and their properties are briefly demonstrated.

3. Alternative Collinearity Diagnostics

In section 2, it was mentioned that Carrillo-Gamboa and Gunst's suggested diagnostic ($\tilde{\Gamma}_{xx}$) involved some problem. Two alternative collinearity diagnostics that modify $\tilde{\Gamma}_{xx}$ given in (2.5) are introduced in this section. Modifying procedures are concentrated mainly on the positive definiteness of estimated Γ_{xx}.

i) $\tilde{\Gamma}_{xx}^{(1)} = \begin{cases} M_{xx} - \Sigma_{uu}, & \text{if } \hat{\lambda} \geq 1 + n^{-1}, \\ M_{xx} - (\hat{\lambda} - n^{-1}) \Sigma_{uu}, & \text{if } \hat{\lambda} < 1 + n^{-1}, \end{cases}$ where $\hat{\lambda}$ is defined in (2.3).

Above estimator $\tilde{\Gamma}_{xx}^{(1)}$ is adapted from the equation-error model results. That is, $\tilde{\Gamma}_{xx}^{(1)}$ is an modified estimator of Γ_{xx} in equation-error model and it is always positive definite(Fuller 1987, p.172). Furthermore, we have the following useful result(Fuller 1987, p.128):

$$\hat{\lambda} \xrightarrow{p} 1. \tag{3.1}$$

By (3.1), $\tilde{\Gamma}_{xx}^{(1)}$ is a consistent estimator of Γ_{xx} since $\tilde{\Gamma}_{xx}$ is(Carrillo-Gamboa and Gunst, section 2).

ii) $\tilde{\Gamma}_{xx}^{(2)} = M_{xx} - \hat{\lambda}\, \Sigma_{uu}$, where $\hat{\lambda}$ is defined in (2.3).

$\tilde{\Gamma}_{xx}^{(2)}$ is the inverse of a first portion of (2.3). It is not a method of moments estimator of Γ_{xx} in our assumed model. It is tried as the second alternative diagnostic since it is always positive definite and is a consistent estimator of Γ_{xx} by (3.1).

4. Simulation Scheme

To compare the small-sample properties of two suggested alternative diagnostics, the following simulation with 3(that is, $k = 3$) highly correlated error-free predictors is performed. That is, 100 $\pi_{1i} \sim \pi_{3i}$'s are simulated from $MVN(\mu, \Sigma_{xx})$ using subroutine RNMVN of IMSL, where

$$\mu^t = (5.0 \quad 10.0 \quad 15.0) \text{ and } \Sigma_{xx} = \begin{pmatrix} 25.00 & 24.85 & 24.50 \\ & 25.00 & 24.60 \\ & & 25.00 \end{pmatrix}. \tag{4.1}$$

Since we are assuming a functional no-equation-error model, 100 $\pi_{1i} \sim \pi_{3i}$'s simulated above are used as fixed error-free predictors through the whole simulation. Error-free response variable ϕ_i is obtained from the equation $\phi_i = 15.0 + \pi_{1i} + \pi_{2i} + \pi_{3i}$. From the simulated 100 $\pi_{1i} \sim \pi_{3i}$'s, it is found that the smallest eigenvalue of true second-order moment matrix of error-free predictors is 0.00113. The true smallest eigenvalue is supposed to be compared with the smallest eigenvalues of suggested estimator matrices. The observed

response and predictor variables vector z_i is obtained by generating measurement error vector w_i using subroutine RNMVN again and suming it to error-free ξ_i. Since we are assuming w_i's are iid $MVN(0, \Sigma_{ww})$, we need to specify Σ_{ww} to generate them. Four sets of Σ_{ww} used in this simulation are provided in the below:

$$\text{I}\quad \begin{pmatrix} 12 & 0 & 0 & 0 \\ & 12 & 0 & 0 \\ & & 12 & 0 \\ & & & 12 \end{pmatrix} \qquad \text{II}\quad \begin{pmatrix} 8 & 0 & 0 & 0 \\ & 8 & 0 & 0 \\ & & 8 & 0 \\ & & & 8 \end{pmatrix} \qquad \text{III}\quad \begin{pmatrix} 8 & 4 & 4 & 4 \\ & 8 & 4 & 4 \\ & & 8 & 4 \\ & & & 8 \end{pmatrix} \qquad \text{IV}\quad \begin{pmatrix} 3 & 0 & 0 & 0 \\ & 3 & 0 & 0 \\ & & 3 & 0 \\ & & & 3 \end{pmatrix}.$$

Two cases are considered:

i) Σ_{ww} is completely known.

Four sets of Σ_{ww} given in the above are used in calculating $\tilde{\Gamma}_{\pi\pi}$ and two alternative diagnostics suggested in this work.

ii) Σ_{ww} is completely unknown.

In this case, we need a reasonable estimator of Σ_{ww} since all of suggested collinearity diagnostics include Σ_{uu}. Hence, replicated observations are essential. In this simulation, three replications are applied and from these, an unbiased estimator of Σ_{ww} is obtained(Fuller 1987, p.131). S_{ww} is used to denote an unbiased estimator of Σ_{ww} and in calculating collinearity diagnostics.

Three sample sizes $n = 100$, 500 and 800 are used. For $n = 500$ and 800, values of 100 π_i's which are generated from (4.1) are duplicated 5 and 8 times respectively. Simulation is performed 200 times for each combination of Σ_{ww} and n. Simulation results obtained with the assumption of completely known Σ_{ww} are given in Table 1. Results of the other case are shown in Table 2. Both tables include the mean of 200 smallest root ($\hat{\lambda}$) of three estimators of $\Gamma_{\pi\pi}$ introduced in section 3. Numbers in the parentheses of tables denote the counts how many times negative smallest eigenvalues are obtained among 200.

5. Concluding Remarks

As a criterion for determining the best collinearity diagnostic, the difference between the mean of 200 smallest eigenvalues and the true smallest eigenvalue(=0.00113) is used in this work. Examination of tables reveals that results of completely known Σ_{uu} case are better than those of estimated Σ_{uu} in most cases, which is a reasonable consequence. As n increases and as the magnitude of measurement error variances decreases, the simulation results become better in all cases which is an another expected consequence.

From the results given in tables, it seems that $\widetilde{\Gamma}_{xx}$ is not suitable for collinearity diagnostic since, as was mentioned in Carrillo-Gamboa and Gunst, it gives negative mean of smallest eigenvalues for most combination of Σ_{uu} and n. From the comparison of remaining ones, we conclude that $\widetilde{\Gamma}_{xx}^{(2)}$ is superior to $\widetilde{\Gamma}_{xx}^{(1)}$ in all cases, since the mean of the smallest eigenvalues of $\widetilde{\Gamma}_{xx}^{(2)}$ is closer to 0.00113 than that of $\widetilde{\Gamma}_{xx}^{(1)}$. Hence, as a conclusion, it is recommended to use the smallest eigenvalue of $\widetilde{\Gamma}_{xx}^{(2)}$ as a collinearity diagnostic if it is suspected that linear MEM involves severe collinearity problems.

References

[1] Amemiya, Y. (1985). What should be done when an estimated between-group covariance matrix is not nonnegative definite, *The American Statistician,* Vol. 39, 112-117.

[2] Belsley, D. A., Kuh, E. and Welsch, R. E. (1980). *Regression Diagnostics: Identifying Influential Data and Sources of Multicollinearity*, New York: Wiley and Sons, Inc. .

[3] Carrillo-Gamboa O. and Gunst R. F. (1992). Measurement-Error-Model Collinearities, *Technometrics,* Vol. 34, No. 4, 454-464.

[4] Fuller, W. A. (1987). *Measurement error models*, New York: Wiley and Sons, Inc.

[5] Gunst, R. F. (1983). Regression analysis with multicollinear predictor variables: Definition, detection, and effects, *Communications in Statistics-Theory and Methods,* Vol. 12, 2217-2260.

[6] Jagpal, H. S. (1982). Multicollinearity in structural equation models with unobservable variables, *Journal of Marketing Research,* Vol. 19, 431-439.

[7] Silvey, S. D. (1969). Multicollinearity and Imprecise Estimation, *Journal of the Royal Statistical Society,* Series B, Vol. 31, 539-552.

Table 1. With Completely Known Error Covariance Matrix, Σ_{uw}

Parameter Set	Estimator	$\hat{\lambda}$		
		$n = 100$	$n = 500$	$n = 800$
I	$\tilde{\Gamma}_{xx}$	-.01116 (25)	-.00372 (44)	-.00212 (61)
	$\tilde{\Gamma}_{xx}^{(1)}$	.00600	.00335	.00275
	$\tilde{\Gamma}_{xx}^{(2)}$	.00527	.00316	.00258
II	$\tilde{\Gamma}_{xx}$	-.00719 (40)	-.00204 (57)	-.00083 (83)
	$\tilde{\Gamma}_{xx}^{(1)}$	.00459	.00244	.00243
	$\tilde{\Gamma}_{xx}^{(2)}$	.00404	.00232	.00230
III	$\tilde{\Gamma}_{xx}$	-.00275 (54)	-.00020 (98)	.00007 (118)
	$\tilde{\Gamma}_{xx}^{(1)}$	.00284	.00172	.00149
	$\tilde{\Gamma}_{xx}^{(2)}$	.00259	.00161	.00139
IV	$\tilde{\Gamma}_{xx}$	-.00140 (61)	.00019 (114)	.00051 (143)
	$\tilde{\Gamma}_{xx}^{(1)}$	.00235	.00157	.00146
	$\tilde{\Gamma}_{xx}^{(2)}$	.00212	.00147	.00137

Table 2. With Estimated Error Covariance Matrix, S_{uw}

Parameter Set	Estimator	$\hat{\lambda}$		
		$n = 100$	$n = 500$	$n = 800$
I	$\tilde{\Gamma}_{xx}$	-.01498 (17)	-.00399 (48)	-.00351 (40)
	$\tilde{\Gamma}_{xx}^{(1)}$	.00681	.00394	.00300
	$\tilde{\Gamma}_{xx}^{(2)}$	.00605	.00377	.00288
II	$\tilde{\Gamma}_{xx}$	-.00789 (36)	-.00285 (45)	-.00166 (67)
	$\tilde{\Gamma}_{xx}^{(1)}$	.00564	.00276	.00226
	$\tilde{\Gamma}_{xx}^{(2)}$	.00515	.00265	.00213
III	$\tilde{\Gamma}_{xx}$	-.00348 (54)	-.00027 (97)	-.00007 (109)
	$\tilde{\Gamma}_{xx}^{(1)}$	.00309	.00193	.00168
	$\tilde{\Gamma}_{xx}^{(2)}$	.00282	.00180	.00158
IV	$\tilde{\Gamma}_{xx}$	-.00285 (46)	-.00011 (99)	.00053 (132)
	$\tilde{\Gamma}_{xx}^{(1)}$	.00242	.00152	.00149
	$\tilde{\Gamma}_{xx}^{(2)}$	.00217	.00142	.00135