Analysis of the Prediction Problem in Linear Regression When the Independent Variables Are Measured With Error†

Jai Hyun Byun (Industrial Engineering,
Gyeongsang National University)

Bong Jin Yum (Industrial Engineering, KAIST)

ABSTRACT

In a regression relationship the independent variables are frequently measured with error when measurements are made in the field under less controlled conditions, or when accurate instruments are not available. This paper deals with the prediction problem for the above situation. The integrated mean square error of prediction (IMSE) is developed as a measure of the effect of the errors in the independent variables on the predicted values. The IMSE may be used for assessing the severeness of measurement errors as well as for comparing competing estimators. An example from the area of work measurement is analyzed.

1. Introduction

The classical theory of regression assumes that the independent variables are measured without error. In practice, however, this assumption is frequently violated

---

† This is a summary of the paper to be published in IIE Transactions.

due to experimental and observational errors. For instance, in developing a standard data system for work measurement, the normal time for an activity element of a job may depend on such job characteristics as weight, distance moved, etc. Then, a regression relationship between the normal time and these characteristics needs to be developed to eventually establish a total task time for an existing or a newly proposed job. However, it is often the case in practice that job characteristics cannot be measured exactly, especially when measurements are made in the field under less controlled conditions, or when accurate instruments are not available. Then, a question arises as to how such measurement errors in job characteristics affect the estimated relationship(between the normal time and job characteristics) and the predicted normal time for an activity element in a future job.

An appropriate model for dealing with such cases is the so called errors-in-variables model(EVM), which is further classified into functional and structural one if the variables involved are fixed and random, respectively(Kendall and Stuart[5]). The problem of estimating unknown parameters in the EVM has been extensively discussed in the literature (e.g., see Madansky[7], Moran[8], and Kendall and Stuart[5]). However, the prediction problem has received rather limited attention despite its importance in practice (e.g., see Denton and Kuiper[1], and Hodges and Moore[4] who pointed out the need for such a study). Several studies exist on the prediction problem in the EVM context(see Lindley[6] and Ganse et al.[2]). Recently, Yum and Neuhardt[11] considered a prediction problem for a simple functional relationship model with replicated observations. They compared the relative performance of the ordinary and grouping least squares estimation methods in terms of the integrated mean square error of prediction (IMSE).

The purpose of this paper is first to define the prediction problems for a multiple functional relationship model, and then to provide corresponding analysis methods for assessing the effect of errors in the variables on prediction accuracy as well as for comparing competing estimators.

2. Model and Assumptions

Assume that variables $\xi_1, \xi_2, \ldots, \xi_p$ and η are linearly related as

$$\eta = \beta_1\xi_1 + \beta_2\xi_2 + \ldots + \beta_p\xi_p = \beta'\xi \qquad (1)$$

where $\beta' = (\beta_1, \beta_2, \ldots, \beta_p)$ is a vector of unknown parameters, and $\xi' = (\xi_1, \xi_2, \ldots, \xi_p)$. In an experiment to estimate the relationship suppose one observes

$$\begin{aligned} y &= \eta + v \\ x &= \xi + u \end{aligned} \qquad (2)$$

where u is a (p×1) vector of random measurement errors in x while v may be interpreted as a natural(or inherent system) variation or a random measurement(or recording) error in y. We further assume that v and u are distributed as

$$\begin{pmatrix} v \\ u \end{pmatrix} \sim \text{MVN} \left\{ 0, \sum = \begin{pmatrix} \sigma_v^2 & 0 \\ 0 & \Sigma_u \end{pmatrix} \right\}. \qquad (3)$$

In terms of the work measurement example, ξ_i may be interpreted as the true value of the i-th job characteristic, η as the expected normal time of an activity element, u_i as a random measurement error in the i-th job characteristic, and v as a natural variation of the normal time around its expected value η.

Suppose ξ in Eq.(1) is a vector, each of its components is a fixed variable. That is, assume that Eq.(1) represents a functional relationship between η and ξ. Then,

$$\begin{pmatrix} y \\ x \end{pmatrix} \sim \text{MVN} \left\{ \begin{pmatrix} \beta'\xi \\ \xi \end{pmatrix}, \Sigma = \begin{pmatrix} \sigma_v^2 & 0 \\ 0 & \Sigma_u \end{pmatrix} \right\}. \qquad (4)$$

Therfore,

$$E(y|x) = \beta'\xi = \eta, \qquad (5)$$

$$V(y|x) = \sigma_v^2. \qquad (6)$$

As can be noted in Eq.(5) the regression of y on x involves unknown ξ, which must be estimated. The "best" predictor of y given x is $\beta'\xi$. An estimate of the best predictor is then given by

$$y = b'x \qquad (7)$$

where b is an estimator of β.

3. The Integrated Mean Square Error of Prediction

Suppose we have n independent observations $\{(y_i, x_i), i = 1, 2, \ldots, n\}$ which satisfy the conditions in Eqs.(1), (2), and (3). The unknown β may be estimated by some selected method. For instance, the ordinary least squares (OLS) estimation yields

$$b_{OLS} = (X'X)^{-1}X'y \qquad (8)$$

where

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}, \tag{9}$$

$$y' = (y_1, y_2, \ldots, y_n). \tag{10}$$

In case where Σ_u is known, the so called corrected least squares (CLS) estimation gives (Schneeweiβ[9])

$$b_{CLS} = (X'X - n\Sigma_u)^{-1}X'y. \tag{11}$$

When Σ is known (or known to within a proportionality factor), the maximum likelihood(ML) estimator of β can be obtained from Gleser[3] with some modification.

Assuming that the prediction is made over the entire region of R where $\xi \in R$, we are interested in some "average" behavior of the predicted values. Adopted as a criterion is the integrated mean square error of prediction(IMSE) which is defined by

$$IMSE = \int_R MSE(\hat{y}_f)w(\xi_f)d\xi_f. \tag{12}$$

The weight function $w(\xi_f)$ describes a priori assumptions on the relative importance of ξ_f values and satisfies

$$\int_R w(\xi_f)d\xi_f = 1. \tag{13}$$

We further assume that

$$\int_R \xi_f\xi_f'w(\xi_f)d\xi_f = M \tag{14}$$

exists. We then have

$$\text{IMSE} = \text{tr}[\{V + (\beta + \phi)(\beta + \phi)'\}\Sigma_u] + \text{tr}\{(V + \phi\phi')M\} + \sigma_v^2$$

$$= \text{tr}\{(V + \phi\phi')(\Sigma_u + M)\} + (\sigma_v^2 + \beta'\Sigma_u\beta) + 2\beta'\Sigma_u\phi. \tag{15}$$

where

$$V = \text{Cov}(b), \tag{16}$$

$$\phi = E(b) - \beta. \tag{17}$$

Estimators can be compared in terms of IMSE. In section 4, three(OLS, CLS, ML) estimators are compared for a work measurement example.

4. An Example

Smith [10] illustrates an example in which six activity elements are identified in a job family of horizontal boring mill operations. Especially, it is found that the fifth element "hoist and aside" depends on such job characteristics as "weight" of a work piece and "distance moved" from the mill to a storage area. A series of time studies were conducted, and then using regression analysis a predictive equation was developed between the normal time(dependent variable) for "hoist and aside" and the two job characteristics(independent variables). Now suppose that "distance moved" and "weight" are grossly measured by imprecise instruments. Such measuring practices may not be unusual in the field where accurate instruments are not available and taking exact measurements is time-consuming and costly.

For the current example we are interested in determining how various magnitudes of errors in the job characteristics affect the average behavior of the predicted normal time for "hoist and aside".

Given a set of parameter values approximate IMSE's for the OLS, CLS, and ML estimation methods can be calculated. For various combinations of σ_1(S.D. of the error in

measuring "distance moved") and σ_2(S.D. of the error in measuring "weight"), Table 1 summarizes values of $\delta = 100(IMSE - IMSE_0)/IMSE_0$ where $IMSE_0$ is the IMSE when the independent variables("distanced moved" and "weight") are measured without error.

Table 1. Percent Increse in IMSE For the Example

σ_2 \ σ_1	0	8	25	40	80
0	0^1	.1004	.9807	2.512	10.06
	0^2	.1002	.9778	2.499	9.919
	0^3	-	-	-	-
0.1	1.080	1.180	2.061	3.592	11.15
	1.079	1.180	2.057	3.579	11.10
	-	1.180	2.061	3.592	11.13
0.3	9.724	9.825	10.71	12.24	19.81
	9.697	9.797	10.68	12.20	19.64
	-	9.822	10.70	12.24	19.79
0.5	27.04	27.14	28.02	29.56	37.15
	26.84	26.94	27.83	29.36	36.82
	-	27.11	28.00	29.54	37.13
1.0	108.64	108.74	109.64	111.20	118.90
	105.68	105.78	106.68	108.23	115.82
	-	108.31	109.21	110.78	118.51

1 : CLS
2 : OLS
3 : ML
4 : Not Available

From Table 1 we observe the following.

1. The three estimation methods preform similarly in terms of δ, although the OLS method becomes slightly better than the others as σ_2 increases.
2. The percent increase in IMSE is more sensitive to the change in σ_1 than in σ_2.
3. A substantial amount of increase in IMSE may occur depending on σ_1.

5. Conclusions

The IMSE is suggested as a measure of overall, average prediction accuracy when the independent variables are subject to error. The present analysis may provide an indication of the severeness of measurement errors by comparing $IMSE_0$(IMSE when there is no error in the independent variables) and IMSE. The sensitivity analysis of IMSE with respect to the measurement error in each x_i(the i-th independent variable) is also investigated. Such analysis is useful for determining which measurement error is more responsible for the increase in IMSE and ultimately for presenting a guideline as to which one should be controlled among others. Developing a criterion for assessing the severeness of measurement errors in each independent variable needs further investigation.

REFERENCES

1. Denton,F.T.and Kuiper, J., "The Effect of Measurement Errors on Parameter Estimates and Forecasts: A Case Study Based on the Canadian Preliminary National Accounts." Rev. Econ. Statist., Vol.47, pp.198-206, 1965.

2. Ganse,R.A., Amemiya,Y., and Fuller,W.A., "Prediction

'When Both Variables Are Subject to Error, with Application toEarthquake Magnitudes." J.Amer. Statist. Ass., Vol.78, pp.761-765, 1983.

3. Gleser,L.J., "Estimation in a Multivariate Errors in Variables Regression Model : Large Sample Results." Ann. Statist., Vol.9, pp.24-44, 1981.

4. Hodges,S.D. and Moore,P.G., "Data Uncertainties and Least Squares Regression." Appl. Statist., Vol.21, pp.185-195, 1972.

5. Kendall,M.G. and Stuart,A., The Advanced Theory of Statistics, Vol.2, 4th Ed., Macmillan, New York, 1979.

6. Lindley,D.V., "Regression Lines and the Linear Functional Relationship." J. R. Statist. Soc., Supp., Vol.9, pp.219-244, 1947.

7. Madansky,A., "The Fitting of Straight Lines When Both Variables Are Subject to Error." J. Amer. Statist. Ass., Vol.54, pp.173-205, 1959.

8. Moran,P.A.P, "Estimating Structural and Functional Relationships." J. Multivariate Anal., Vol.1, pp.232-255, 1971.

9. Schneeweiβ,H., "Consistent estimation of a Regression with Errors in the Variables." Metrika, Vol.23, pp.101-115, 1976.

10. Smith,G.L.,Jr., Work Measurement: A Systems Approach, Grid Publishing, Inc., Columbus, 1978.

11. Yum,B.J. and Neuhardt,J.B., "Analysis of the Prediction Problem in a Simple Functional Relationship Model." IIE Trans., Vol.16, pp.177-184, 1984.