Journal of the Korean Institute
of Industrial Engineers
Vol. 19, No.1, Mar., 1993

독립변수의 측정오차가 예측에 미치는 영향을 평가하기 위한 기준개발+

Development of a Criterion for Assessing the Influence of the Measurement Errors in the Independent Variables on Prediction+

변 재 현*

Jai-Hyun Byun*

Abstract

In developing a multiple regression relationship, independent variables are frequently measured with error. For these situations the problem of estimating unknown parameters has been extensively discussed in the literature while little attention has been given to the prediction problem. In this paper a criterion is developed for assessing the severeness of mesurement errors in each independent variable on the predicted values. Using the developed criterion we can present a guideline as to which measurement error should be controlled for a more accurate prediction. Proposed methods are illustrated with a standard data system in work measurement.

1. 序 論

공학이나 자연과학의 여러분야에서 變數들간의 相互關聯性을 파악하기 위하여 回歸分析을 자주 하게 되는데, 이 때 독립변수에 오차가 포함되어 측정될 경우가 많다. 특히, 환경의 변화, 취급의 부주의, 노후등으로 인하여 測定計器가 부정확하거나 제어상태가 충실하지 못한 생산현상에서는 정확한 측정을 할 수 없기 때문에 독립변수의 측정에 오차가 수반된다.

예를 들면, 作業測定분야에서 표준자료시스템을 개발하는데 있어서 어떤 作業要素의 正味時間値가 두개의 作業特性(예를 들면, 작업대상물의 무게와 움직인 거리)에 따라서 변하는 경우, 요소작업의 정미시간치를 종속변수로 놓고 작업특성을 독립변수로 하여 回歸分析을 이용해서 豫測方程式을 구한다. 그런 다음, 새로운 작업의 작업특성치가 주어지면 추정된 예측방정식으로 부터 작업요소의 정미시간치를 豫測할 수 있다. 그러나, 흔히 정확한 측정이 제대로 이루어지지 않는 生産現場에서는 작업특성치가 誤差를 수반하여 측정된다. 이럴 때에는 미래의 새로운 作業에 대한 作業要素의 正味時間을 豫測하기 위해 예측시 각 作業特性(獨立變數)을 測定하는데 있어서 오차가 豫測値의 正確性에 미치는 영향이 각각 어느 정도인가를 파악하여 예측의 정확성에 큰 영향을 미치는 작업특성은 집중적으로 統制할 필요가 있다.

+이 논문은 1990년도 문교부지원 학술진흥재단의 자방대육성 학술연구조성비에 의해 연구되었음

* 경상대학교 산업공학과

위와 같이 獨立變數에 測定誤差가 있을때 回歸模型을 變數誤差模型(Errors-in-Variables Model, EVM)이라고 하는데, 변수오차모형에 관하여 母數推定에 관한 연구는 많이 있다(Fuller[2], Kendall과 Stuart[5], Moran[7]). 하지만 변수오차모형의 豫測에 관한 문제는 그 실용적 중요성에도 불구하고(Hodges와 Moore[4]) 별로 관심을 받지 못했다. 변수오차모형의 예측에 관한 기존의 연구를 살펴보면, 우선 Lindley[6]는 독립변수의 觀測値가 주어졌을 때 종속변수의 最尤(most likely) 예측치가 존재하기 위한 조건을 확립했으며 Ganse 등[3]은 Lindley의 연구결과를 추정모집단과 예측모집단이 서로 다른 경우로 확장하였다. Yum과 Neuhardt[12]는 반복이 있는 單純變數誤差模型(Simple EVM)에 대하여 예측의 통합평균제곱오차(Integrated Mean Square Error of Prediction, IMSE)를 개발하고 IMSE의 관점에서 보통최소제곱추정(Ordinary Least Square Estimation)방법과 그룹 최소제곱추정(Grouping Least Square Estimation)방법을 비교하였다. 최근에는 Yum과 Byun[11]이 多重變數誤差模型(Multiple EVM)에 대한 IMSE를 개발하여 각 독립변수의 측정오차에 따른 IMSE의 상대적 크기를 예제에 통하여 비교하였으나 豫測時 각각의 測定誤差가 豫測의 正確性에 미치는 基準은 개발하지 못했다.

本 硏究의 目的은 獨立變數에 測定誤差가 있을 때 回歸模型인 多重變數誤差模型에서 關係式의 推定 후, 豫測時에 각 獨立變數의 誤差가 豫測의 正確性에 미치는 영향을 평가하기 위한 기준을 개발하는데에 있다.

本 論文은 다음과 같이 構成되어 있다. 제2절에서는 多重變數誤差模型의 推定과 豫測을 위한 模型을 소개하였고, 3절에서는 평균제곱오차로 부터 통합평균제곱오차의 형태 및 이것의 추정치를 구하는 방법이 제시되었다. 예측시에 어떤 독립변수의 측정오차가 豫測의 正確性에 더 큰 영향을 미치는가를 評價하기 위한 基準은 4절에서 밝혀지며, 5절에서는 標準資料의 예를 통하여 4절에서 개발된 기준이 제대로 적용되었는지의 여부를 평가하였다.

2. 推定과 豫測에 관한 模型

일반적인 다중회귀모형은 다음과 같이 표현된다.

$$y=\beta_0+\beta_1\xi_1+\beta_2\xi_2+\cdots+\beta_p\xi_p+v \qquad (2.1)$$

여기서 y는 종속변수, β_0, β_1, …, β_p는 미지의 모수, ξ_1, ξ_2, …, ξ_p는 독립변수, 그리고 v는 실험에 고유한 변동이다. 그런데 독립변수가 오차를 수반하는 경우에, 우리가 관측하는 값은 ξ_i가 아니고

$$x_i=\xi_i+u_i,\ i=1,\ 2,\ \cdots,\ p \qquad (2.2)$$

이다. 여기서 u_i는 i번째 독립변수의 오차이다. 관계식 (2.1)을 추정하기 위해 n개의 표본을 취하여 ξ_{ij}, $y_j(j=1,\ 2,\ \cdots,\ n)$를 측정한다고 할 때 ξ_{ij}의 측정에는 측정오차 u_{ij}가 수반되어 x_{ij}가 관측되므로 다음과 같은 모형을 생각할 수 있다.

$$y_j=\beta_0+\beta_1\xi_{1j}+\beta_2\xi_{2j}+\cdots+\beta_p\xi_{pj}+v_j \qquad (2.3)$$
$$x_{ij}=\xi_{ij}+u_{ij}.$$

단, $i=1,\ 2,\ \cdots,\ p$; $j=1,\ 2,\ \cdots,\ n$.

j번째 측정치에 대하여 독립변수의 오차벡터를 다음과 같이 정의하자.

$$u_j=(0,\ u_{1j},\ u_{2j},\ \cdots,\ u_{pj})'.$$

첫번째 요소를 0으로 한 것은 식(2.3)의 우변의 첫째항 β_0를 $\beta_0\xi_{0j}$, $\xi_{0j}\equiv 1$로 보고 ξ_{0j}의 오차 u_{0j}를 0으로 간주했기 때문이다. 그리고, 벡터$(v_j,\ u_j')'$는 평균이 0벡터이고 다음과 같은 共分散 行列(covariance matrix)을 갖는다고 가정한다.

$$\mathrm{Cov}(v_j,\ u_j')'=\begin{bmatrix}\delta^2 & 0\\ 0 & \Sigma\end{bmatrix}$$

단,

$\delta^2=\mathrm{Var}(v_j)$

$\Sigma=\mathrm{Cov}(u_j)=\mathrm{diag}(0,\ \sigma^2_1,\ \sigma^2_2,\ \cdots,\ \sigma^2_p)$.

$\sigma^2_i=\mathrm{Var}(u_{ij}),\ i=1,\ 2,\ \cdots,\ p.$

관계식의 추정을 위해 독립변수와 종속변수의 관측치에 대한 행렬과 벡터를 다음과 같이 정의한다.

$$X=\begin{bmatrix} 1 & x_{11} & x_{21} & \cdots & x_{p1} \\ 1 & x_{12} & x_{22} & \cdots & x_{p2} \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ 1 & x_{1n} & x_{2n} & \cdots & x_{pn} \end{bmatrix}$$

$$y=(y_1, y_2, \cdots, y_n)'$$

아울러 관계식 (2.1) 또는 (2.3)에서 미지의 모수들로 이루어진 벡터 β를 다음과 같이 정의한다.

$$\beta=(\beta_0, \beta_1, \beta_2, \cdots, \beta_p)'.$$

그러면, β에 대한 보통최소제곱(Ordinary Least Squares, OLS) 추정치 b는 다음과 같이 주어진다.

$$b=(X'X)^{-1}X'y$$

지금까지 模型과 推定方法에 대해 살펴 보았는데, 이제 豫測을 위한 模型에 관하여 살펴보자. 미래의 종속변수와 독립변수의 참값 사이에 다음과 같은 관계가 성립한다고 하자. (식 (2.1) 참조)

$$y_f=\beta_0+\beta_1\xi_{1f}+\beta_2\xi_{2f}+\cdots+\beta_p\xi_{pf}+v_f \quad (2.4)$$

그리고 $\xi_{if}(i=1, 2, \cdots, p)$를 측정함에 있어서 추정실험에서와 마찬가지로 다음과 같이 측정오차 u_{if}가 수반된다고 가정한다.

$$x_{if}=\xi_{if}+u_{if}.$$

v_f의 평균은 0. 분산은 δ^2이고, $u_f=(0, u_{1f}, u_{2f}, \cdots, u_{pf})'$는 평균 0, 분산 $\sum_f=\mathrm{diag}(0, \sigma^2_{1f}, \sigma^2_{2f}, \cdots, \sigma^2_{pf})$을 가지며, v_f와 u_f는 서로 상관관계가 없다고 가정한다. 여기서 $u_{if}(i=1, 2, \cdots, p)$의 분산 σ^2_{if}은 추정실험에서의 분산 σ^2_i과는 일반적으로 다르다는 것에 주의할 필요가 있다.

식 (2.4)로 부터 y_f의 최선의 예측치는 $\beta'\xi_f$, $\xi_f=(1, \xi_{1f}, \xi_{2f}, \cdots, \xi_{pf})$이나, β와 ξ_f는 알 수 없으므로 y_f의 예측치로서

$$y_f=b'x_f \quad (2.5)$$

로 삼는다. 식 (2.5)에서 b는 추정실험을 통해 구한 β의 OLS 추정 벡터이며,

$$x_f=(1, x_{1f}, x_{2f}, \cdots, x_{pf})'$$

이다.

3. 예측의 통합평균제곱오차

식 (2.5)의 $\hat{y}_f$가 y_f에 얼마나 가까운가를 나타내는 尺度로서 평균제곱오차(mean square error, MSE)를 고려하고자 한다. 또한, 미래의 豫測値들이 여러개 있을 때 그들의 平均的 行態를 나타내는 척도로서 소위 통합평균제곱오차 IMSE를 채택하고자 한다.

IMSE를 구하기 전에 우선 x_f에 대한 $\hat{y}_f$의 조건부 MSE를 구해보면

$$\mathrm{MSE}(\hat{y}_f \mid x_f)$$
$$=E\{(\hat{y}_{if}-y_{if})^2 \mid x_f\}$$

가 된다. x_f에 대하여 조건부 평균제곱오차의 기대값을 취하면 $\hat{y}_f$의 평균제곱오차를 얻을 수 있는데, 그 전에 다음을 정의한다.

$$V=\mathrm{Cov}(b)$$
$$\phi=E(b)-\beta$$

Seber[9]의 Theorem 1.7을 이용하여 수식을 정리하면, $\hat{y}_f$의 MSE는 다음과 같이 된다.

$$\mathrm{MSE}(\hat{y}_f)=E[\mathrm{MSE}(\hat{y}_f \mid x_f)]$$
$$=\mathrm{tr}[\{V+(\beta+\phi)(\beta+\phi)'\}\sum_f]$$
$$+\xi_f'(V+\phi\phi')\xi_f+\delta^2 \quad (3.1)$$

예측이 어떤 R이라는 흥미영역(Region of Interest)에서 이루어진다면, 우리는 예측치들의 "평균적" 행태에 관심을 갖게 된다. 독립변수의 측정오차가 예측에 미치는 영향에 대한 尺度로서 IMSE를 다음과 같이 정의한다.

$$\mathrm{IMSE}=\int_R \mathrm{MSE}(\hat{y}_f)w(\xi_f)d\xi_f. \quad (3.2)$$

여기서 加重係數(weight function) $w(\xi_f)$는 ξ_f값들의 상대적 중요성(예를 들어, 각 ξ_f값을 갖는 작업들이 출현할 빈도)을 나타내며 다음을 만족한다.

$$\int_R w(\xi_f)d\xi_f=1. \quad (3.3)$$

그리고 다음과 같이 ξ_f의 2차 모멘트가 존재한다고 가정한다.

$$\int_R \xi_f\xi_f' \; w(\xi_f)d\xi_f=M. \quad (3.4)$$

식 (3.1)-(3.4)를 조합하면 다음과 같이 IMSE를 구할 수 있다.

$$IMSE=tr[\{V+(\beta+\phi)(\beta+\phi)'\}\Sigma_f] +tr\{(V+\phi\phi')M\}+\delta^2 \quad (3.5)$$

만약에 예측시에 독립변수의 측정오차가 없다면, 식 (3.5)에서 $\Sigma_f=0$이므로 IMSE는 다음과 같이 표현된다.

$$IMSE_0=tr\{(V+\phi\phi')M\}+\delta^2$$

식 (3.5)의 IMSE를 추정하기전에 우선 독립변수의 관측치 참값에 대한 행렬을 다음과 같이 정의한다.

$$\Xi=\begin{bmatrix} 1 & \xi_{11} & \xi_{21} & \cdots & \xi_{n1} \\ 1 & \xi_{12} & \xi_{22} & \cdots & \xi_{n2} \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ 1 & \xi_{1n} & \xi_{2n} & \cdots & \xi_{nn} \end{bmatrix}$$

IMSE의 推定을 위해서는 식(3.5)의 未知의 값들을 推定해야 하는데, OLS추정법을 이용했을 때 공분산 행렬 V와 편의 벡터 ϕ의 근사값은 Davies와 Hutton[1]으로 부터 다음과 같이 구할 수 있다.

$$\hat{V}\simeq n^{-1}\{\delta^2(M_\xi+\Sigma)^{-1}+(M_\xi+\Sigma)^{-1} T(M_\xi+\Sigma)^{-1}\} \quad (3.6)$$

이며

$$T=(M_\xi+\Sigma)\beta'M_\xi(M_\xi+\Sigma)^{-1}\Sigma(M_\xi+\Sigma)^{-1}M_\xi\beta+\Sigma\beta'M_\xi(M_\xi+\Sigma)^{-1}\Sigma M_\xi^{-1}\Sigma(M_\xi+\Sigma)^{-1}M_\xi\beta -\Sigma(M_\xi+\Sigma)^{-1}M_\xi\beta\beta'M_\xi(M_\xi+\Sigma)^{-1}\Sigma, \quad (3.7)$$

$$M_\xi=\sum_{n\to\infty} n^{-1}(\Xi_n'\Xi_n) : \Xi_n\equiv\Xi. \quad (3.8)$$

그리고,

$$\hat{\phi}\simeq-(M_\xi+\Sigma)^{-1}\Sigma\beta.$$

미지의 모수를 추정하기 위해 Seber[9]로 부터 다음의 결과를 얻을 수 있다.

$$E(X'X/n)=(\Xi'\Xi)/n+\Sigma. \quad (3.9)$$

$$E(S^2)=\delta^2+\beta'\Sigma\beta, \quad (3.10)$$

여기서

$$S^2=(y-Xb)'(y-Xb)/(n-p-1). \quad (3.11)$$

식 (3.8)로 부터 M_ξ는 $(\Xi'\Xi)/n$으로 근사될 수 있으며, 이는 다시 식 (3.9)를 이용하면 $(X'X)/n-\Sigma$로 근사화된다. β는 b로 추정되며, 식 (3.10)으로 부터 δ^2은 $S^2-b'\Sigma b$로 근사화된다. 만일 Σ 또는 Σ_f가 알려지지 않은 경우, 이들은 과거의 자료나 실험을 통하여 추정되어야 하며, 2차 모멘트 행렬 M은 미래에 관심있는 영역에서 식 (3.4)를 이용하여 추정되어야 한다.

4. 比較基準의 開發

3절에서 구한 IMSE는 독립변수의 오차가 예측의 정확성에 전체적으로 어느 만큼 영향을 미치는가를 나타낸다. 이제 예측시 각 독립변수의 측정오차에 따른 IMSE의 敏感度를 分析하여 어떤 독립변수의 測定誤差가 豫測의 正確性에 더 큰 영향을 미치는가를 評價하기 위한 比較基準을 세우고, 결과적으로 영향력이 큰 변수들을 집중적으로 統制함으로써 보다 바람직한 예측치를 얻을 수 있도록 할 것이다.

식 (3.5)의 IMSE를 σ^2_{ii}에 대하여 편미분을 하면,

$$\frac{\partial(IMSE)}{\partial(\sigma^2_{ii})}$$

$$=\frac{\partial}{\partial(\sigma^2_{if})}\left\{\mathrm{tr}[\{V+(\beta+\phi)(\beta+\phi)'\}\Sigma_f]\right\} \tag{4.1}$$

이제

$V_{ij}=\hat{V}$의 (i, j)번째 요소,

$A_{ij}=V_{ij}+(\beta_i+\phi_i)(\beta_j+\phi_j)$,

i=1, 2, …, p : j=1, 2, …, p,

라고 하면,

$$\{V+(\beta+\phi)(\beta+\phi)'\}\Sigma_f$$

$$=\begin{bmatrix} 1 & A_{11} & A_{21} & \cdots & A_{p1} \\ 1 & A_{12} & A_{22} & \cdots & A_{p2} \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ 1 & A_{1n} & A_{2n} & \cdots & A_{pn} \end{bmatrix} \cdot$$

$$\begin{bmatrix} 0 & & & & & \\ 0 & \sigma^2_{1f} & & & & \\ \cdot & \cdot & \cdot & & & \\ \cdot & \cdot & & \cdot & & \\ \cdot & \cdot & & & \cdot & \\ 0 & 0 & \cdot & \cdot & \cdot & \sigma^2_{pf} \end{bmatrix}$$

$$=\begin{bmatrix} 0 & A_{11}\sigma^2_{1f} & A_{21}\sigma^2_{2f} & \cdots & A_{p1}\sigma^2_{pf} \\ 0 & A_{12}\sigma^2_{1f} & A_{22}\sigma^2_{2f} & \cdots & A_{p2}\sigma^2_{pf} \\ \cdot & \cdot & \cdot & \cdots & \\ \cdot & \cdot & \cdot & \cdots & \\ \cdot & \cdot & \cdot & \cdots & \\ 0 & A_{1p}\sigma^2_{1f} & A_{2p}\sigma^2_{2f} & \cdots & A_{pp}\sigma^2_{pf} \end{bmatrix} \tag{4.2}$$

식 (4.2)로 부터,

$$\begin{aligned} &\mathrm{tr}[\{V+(\beta+\phi)(\beta+\phi)'\}\Sigma_f] \\ &=\sum_{i=1}^{p} A_{ii}\sigma^2_{if} \\ &=\sum_{i=1}^{p}[\{V_{ii}+(\beta_i+\phi_i)^2\}\sigma^2_{if}] \end{aligned} \tag{4.3}$$

식 (4.3)을 (4.1)에 대입하면

$$\frac{\partial(\mathrm{IMSE})}{\partial(\sigma^2_{if})}=V_{ii}+(\beta_i+\phi_i)^2 \tag{4.4}$$

그런데, b의 공분산 행렬 V의 추정치인 식 (3.6)의 $\hat{V}$을 보면 그 형태가 다소 복잡하다. Hodges와 Moore[4]는 σ^2_i(i=1, 2, …, p)이 그리 크지 않을 때에는 V의 추정치가 다음과 같이 간략하게 표현될 수 있다는 것을 보였다.

$$\hat{V}\simeq(X'X)^{-1}\{\beta'\Sigma\beta+\delta^2\} \tag{4.5}$$

식 (4.4)와 (4.5)로 부터 σ^2_{if}(i=1, 2, …, p)에 의한 IMSE의 민감도는

$$(X'X)^{-1}_{ii}\{\beta'\Sigma\beta+\delta^2\}+(\beta_i+\phi_i)^2 \tag{4.6}$$

이다. 여기서 $(X'X)^{-1}_{ii}$는 $(X'X)^{-1}$ 행렬의 (i, i)번째 요소이다. 실험데이터가 주어졌을 때에는 식 (4.6)의 $(X'X)^{-1}_{ii}$는 데이터로 부터 직접 구할 수 있고, $\beta'\Sigma\beta+\delta^2$은 식(3.11)의 S^2으로 근사되며, $\beta_i+\phi_i$의 근사치로 b_i를 쓸 수 있다. 실험데이터가 주어지지 않은 상태에서 모수의 값을 알고 있을 때의 민감도 분석을 하기 위해서는, 식(3.9)로 부터 (X'X)대신에 $(\Xi'\Xi+n\Sigma)$를 쓰면 된다.

5. 例題

본 절에서는 4절에서 개발된 예측시 각 독립변수의 측정오차가 IMSE에 미치는 영향을 나타내는 기준이 제대로 적용되는지의 여부를 標準資料의 예를 통하여 알아보고, 전산실험을 하여 각 오차의 여러가지 값에 대한 IMSE의 변화를 관찰하였다.

독립변수의 측정오차가 총작업시간의 예측에 미치는 영향을 Smith[10]의 수평보오링밀링작업(Horizontal Boring Mill Operation)의 예를 통해서 알아보기로 한다. 수평보오링밀링이 끝난 작업물을 荷役하는 活動(unloading activity)은 6개의 作業要所로 構成되어 있는데, 2개의 작업요소("크레인 부름", "크레인 정위치")는 어떤 작업에 대해서든지 그 시간이 일정하고, 다른 4개의 작업요소("체인 걸기", "체인길이 조정", "들어서 옆으로 옮김", 그리고 "체인 제거")는 여러가지 작업에 대하여 그 시간이 다르게 나타났으며, 그 중에서 "들어서 옆으로 옮김"은 작업대상물의 "무게"

와 밀링 머시인에서 저장장소까지의 "이동 거리"에 의존하였다.

이 작업요소의 정미시간치 y는 작업특성 "무게"의 참값 ξ_1과 "이동거리" ξ_2와 다음과 같은 관계를 갖는다.

$$y=\beta_0+\beta_1\xi_1+\beta_2\xi_2+v$$

관계식의 추정을 위해 n개의 작업을 선택하고 각 작업의 정미시간치를 r번 반복하여 관측한다. 관측된 데이터에 대한 표현을 위하여 다음을 정의한다.

ξ_{1k} = k번째 작업의 "무게"의 참값
ξ_{2k} = k번째 작업의 "이동 거리"의 참값
u_{1k} = k번째 작업의 "무게"의 측정오차
v_{2k} = k번째 작업의 "이동 거리"의 측정오차
v_{kq} = k번째 작업의 q번째 반복에 대한 정미시간 추정실험에 고유한 변동

그러면 k=1, 2, ⋯, n ; q=1, 2, ⋯, r에 대하여, 우리는 다음과 같은 관측치를 얻게 된다.

$$\begin{aligned} &y_{kq}=\beta_0+\beta_1\xi_{1k}+\beta_2\xi_{2k}+v_{kq} \qquad (5.1)\\ &\begin{cases} x_{1k}=\xi_{1k}+u_{1k}\\ x_{2k}=\xi_{2k}+u_{2k}\end{cases}\end{aligned}$$

분석의 편의상, 각 k에 대해서 y_{kq}, q=1, 2, ⋯, r의 표본평균을 취하여 그 값을 $\bar{y}_k$라 하면 식 (5.1)은 다음과 같이 쓸 수 있다.

$$\bar{y}_k=\beta_0+\beta_1\xi_{1k}+\beta_2\xi_{2k}+\bar{v}_k$$

단,

$$\bar{y}_k=\sum_{q=1}^{r} y_{kq}/r,$$

$$\bar{v}_k=\sum_{q=1}^{r} v_{kq}/r.$$

2절의 추정실험 모형과 연관지어 보면, 식(2.3)의 y_k와 v_k가 본 예제에서는 각각 $\bar{y}_k$와 $\bar{v}_k$로 대치되었음을 알 수 있다. 아울러 독립변수의 오차에 대해서는

$$u_k=(0,\ u_{1k},\ u_{2k})',$$
$$\Sigma=\mathrm{Cov}(u_k)=\mathrm{diag}(0,\ \sigma^2_1,\ \sigma^2_2)$$

이 된다. 본 예에서는

$$\Sigma=\begin{bmatrix} 0 & & \\ 0 & 1 & \\ 0 & 0 & 50^2 \end{bmatrix}$$

으로 가정한다. y_{kq} 대신에 $\bar{y}_k$를 종속변수로 하여 β_0, β_1, β_2의 OLS 추정량 b_0, b_1, b_2를 구할 수 있다.

미래의 어떤 작업 f에 대하여 우리는 다음과 같은 정규시간을 예측하는 데에 관심을 가지게 된다.

$$y_f=\beta_0+\beta_1\xi_{1f}+\beta_2\xi_{2f}+v_f.$$

y_f의 예측을 위해 작업특성인 무게와 이동거리를 측정하게 되는데, 수리적으로는

$$\begin{cases} x_{1f}=\xi_{1f}+u_{1f}\\ x_{2f}=\xi_{2f}+u_{2f}\end{cases}$$

로 쓸 수 있다.

v_f는 평균 0, 분산 δ^2을 갖고, $u_f=(0,\ u_{1f},\ u_{1f})'$는 v_f와 독립이며, 평균 0과 다음과 같은 공분산 행렬을 갖는다고 가정한다.

$$\Sigma_f=\begin{bmatrix} 0 & & \\ 0 & \sigma^2_{1f} & \\ 0 & 0 & \sigma^2_{2f} \end{bmatrix}$$

정규시간의 예측치 $\hat{y}_f$는

$$\hat{y}_f=b_0+b_1x_{1f}+b_2x_{2f}$$

이 된다. 본 예에서는 모수의 값이 주어진 상태에서 IMSE가 σ^2_{1f}과 σ^2_{2f}에 어느 만큼 민감한가를 보이는데 관심이 있다. 추정을 위한 실험에 선택된 작업특성의 참값은 표 1과 같다고 하자.

표 1. 작업특성의 참값

무게(kg) ξ_1	이동거리(m) ξ_2
350	10
1600	10
1600	20
2700	20
350	30
2700	30

ξ_1과 ξ_2의 각각의 조합에 대하여 15번의 반복측정을 하며, 모수의 값은 $\beta_0=0.2237157$, $\beta_1=0.000097$, $\beta_2=0.025466$, 그리고 $\delta^2=0.0009$라고 하자(Smith[10], p.79참조).

OLS 추정법을 이용했을 때의 IMSE를 계산하기 위해 M_ξ는 표 1의 데이터를 이용하여 계산한 $(\Xi'\Xi)/n$으로 근사치를 구할 수 있는데 그 값은 다음과 같다.

$$M_\xi=\begin{bmatrix} 1 & & \\ .155\times10^{4} & .3324\times10^{7} & \\ .2\times10^{2} & .3283\times10^{5} & .4667\times10^{3} \end{bmatrix}$$

미래에 있어서 독립변수 참값의 2차 모멘트행렬 M이 다음과 같다고 가정한다.

$$M=\begin{bmatrix} 1 & & \\ .15\times10^{4} & .324\times10^{7} & \\ .2\times10^{2} & .324\times10^{5} & .484\times10^{3} \end{bmatrix}$$

위와 같이 모수의 값들이 주어지면 IMSE를 구할 수 있다. 본 예에서는 실험데이터가 주어지지 않은 상태에서 주어진 모수의 값을 이용한 민감도 분석을 하므로 식 (4.6)의 계산을 위해 $(X'X)$대신에 $(\Xi'\Xi+n\sum)$를 이용하고 다른 모수의 값을 계산하여 대입하면 σ^2_{1i}와 σ^2_{2i}에 의한 IMSE의 민감도를 표 2와 같이 구할 수 있다.

표 2. IMSE의 민감도

σ^2_{ii}	IMSE의 민감도
i=1	3.3367×10^{-6}
i=2	9.5267×10^{-4}

표 2로 부터 IMSE가 σ^2_{1i}보다는 σ^2_{2i}에 훨씬 민감하다는 것을 알 수 있다. IMSE의 이러한 민감도를 확인하기 위하여 여러가지 σ_{1i}와 σ_{2i}값에 대한 IMSE의 변화를 SAS/IML[8]을 이용한 프로그램을 작성하여 전산실험을 하였다. 표 3은 여러가지 σ_{1i}와 σ_{2i}의 수준조합에 대하여

$$\Delta=100\cdot(\mathrm{IMSE}-\mathrm{IMSE}_0)/\mathrm{IMSE}_0 \qquad (5.2)$$

의 값을 나타내고 있다. 여기서 IMSE_0는 예측시 독립변수에 오차가 없을 때의 IMSE 값이다. 표3으로부터 IMSE의 증가분은 σ_{2i}의 변화에 아주 민감하며, σ_{1i}는 IMSE의 증가에 거의 영향을 미치지 않는 것을 알 수 있다.

표 3. 수평보오링밀작업 예제에 대한 IMSE의 증가 퍼센트

σ_{2i} \ σ_{1i}	0	10	20	30	40	50
0	0	.0741	.2963	.6667	1.185	1.852
0.2	1.938	2.012	2.234	2.604	3.123	3.790
0.4	7.750	7.825	8.047	8.417	8.936	9.602
0.6	17.44	17.51	17.73	18.11	18.62	19.29
0.8	31.00	31.08	31.30	31.67	32.19	32.85
1.0	48.44	48.51	48.74	49.11	49.63	50.29

즉, σ_{2i}(이동거리를 측정하는데 있어서의 오차의 표준편차)가 σ_{1i}(무게를 측정하는데 있어서의 오차의 표준편차)보다 IMSE를 줄이는 데에 결정적인 역할을 한다는 것을 의미한다. 예를 들어, 현재의 측정체계를 이용한 관계식의 추정에서는 σ_{1i}와 σ_{2i}가 각각 50과 1이라고 하자. 그러면, 보다 나은 예측을 위해서 σ_{1i}를 줄이는 것은 의미가 없고, σ_{2i}를 줄이는 것이 IMSE의 감소를 위해 효과적이다.

6. 結 論

본 硏究는 다중회귀분석에서 獨立變數에 測定誤差가 있을 때, 추정된 회귀관계식을 이용한 豫測時 독립변수의 측정오차가 豫測値들의 正確性에 미치는 影響을 分析 評價하는데에 목적이 있다. 예측시 각 독립변수의 측정오차가 豫測値의 平均的 行態를

나타내는 尺度인 통합평균제곱오차 IMSE에 미치는 상대적 영향을 비교할 수 있도록 각 독립변수의 誤差에 의한 IMSE의 敏感度를 나타내는 基準을 개발하였다. 개발된 기준이 제대로 적용되는지의 여부를 標準資料의 예를 통하여 살펴 보았다. 기준을 이용하면 예측시 어떤 독립변수의 측정오차를 중점적으로 統制해야 보다 나은 예측치를 얻을 수 있는지를 결정할 수 있게 된다. 이 때 물론 각 측정오차의 크기를 줄이는데 필요한 經費, 技術的 問題등이 함께 고려되어야 할 것이다.

본 연구에서는 독립변수의 측정오차의 공분산행렬을 대각행렬로 가정하였는데, 이는 대각행렬이 아닐 경우 식 (4.6)과 같은 기준을 쉽게 구할 수가 없는 난점 때문이었다. 대각행렬이 아닐 때에 기준을 구하는 것은 추후 연구가 필요하다. 또한 독립변수와 종속변수간의 관계가 선형이 아니고 2차적(quadratic)관계일 때의 연구도 추후에 이루어져야 할 것이다.

참 고 문 헌

1. Davies, R.B. and Hutton, B., "The Effect of Errors in the Independent Variables in Linear Regression", *Biometrika*, Vol.11, pp.383-392, 1975.
2. Fuller, W.A., *Measurement Error Models*, John Wiley & Sons, New York, 1987.
3. Ganse, R.A., Amemiya, Y., and Fuller, W. A. "Prediction When Both Variables Are Subject to Error, with Application to Earthquake Magnitudes", *J. Amer. Statist. Assoc.*, Vol.78, pp.761-765, 1983.
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., Ch.29, Griffin, 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. Moran, P.A.P., "Estimating Structural and Functional Relationships", *J. Multivariate Anal.*, Vol.1, pp.232-255, 1971.
8. SAS Institute Inc., *SAS/IML User's Guide*, Release 6.03 Ed., Cary, NC, 1988.
9. Seber, G.A.F., *Linear Regression Analysis*, Wiley, New York, 1977.
10. Smith, G.L., Jr., *Work Measurement: A Systems Approach*, Grid Publishing, Inc., Columbus, 1978.
11. Yum, B.J. and Byun, J.H., "Analysis of the Prediction Problem with Errors in the Variables", *IIE Trans.*, Vol.22, pp.73-83, 1990.
12. 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.