資料誤差와 回歸分析

金　順　基*

I. 序　論

線型模型에 기초를 두는 統計的 推定에서 推定曲線을 구할때 最小自乘法을 사용하는 것은 일반화 되어 있다. 그러나 자료의 측정에 誤差가 발생했을때는 母數의 推定은 어려운 일이다.

일반적인 線型回歸模型 $y=X\beta+u$에는 다음 가정이 있다.

(i) $u\sim N\ (0,\sigma^2 I_n)$, $\sigma^2<\infty$;

(ii) 행렬 X의 階數는 q이다.

最小自乘法에 의하여 구한 母數 β의 推定量은 다음과 같은 성질이 있다.

(1) $\hat{\beta}=(X'X)^{-1}X'y$

(2) $E(\hat{\beta})=\beta+(X'X)^{-1}X'E(u)=\beta$

(3) $V(\hat{\beta})\ =E((\hat{\beta}-\beta)(\hat{\beta}-\beta)')=(X'X)^{-1}\sigma^2$　　[1], [4]

단, y X β는 아래와 같이 표기하기로 한다.

$$(4)\quad y=\begin{bmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ y_n \end{bmatrix},\quad X=\begin{bmatrix} x_{11} & x_{12}\cdots x_{1q} \\ x_{21} & x_{22}\cdots x_{2q} \\ \cdot & \cdot\ \ \cdot \\ x_{n1} & x_{n2}\cdots x_{nq} \end{bmatrix},\quad \beta=\begin{bmatrix} \beta_1 \\ \beta_2 \\ \cdot \\ \cdot \\ \beta_q \end{bmatrix},\quad u=\begin{bmatrix} u_1 \\ u_2 \\ \cdot \\ \cdot \\ u_n \end{bmatrix},\ n>q.$$

본 논문에서는 위와 같은 경우의 線型回歸模型에서 q개의 변수 $X_1, X_2, \ldots, X_q$를 서로 獨立的으로 각각 n회 측정하여 얻은 측정치 $x_{11}, x_{21}, \ldots, x_{n1}; x_{12}, x_{22}, \ldots, x_{n2}, ; \ldots ; x_{1q}, x_{2q}; \ldots, x_{nq}$에 誤差가 발생한다고 가정하여 몇가지의 조건아래서 母數 $\hat{\beta}$를 推定한후 이 推定量의 특성을 검토하기로 한다.

II. 變數의 誤差와 回歸係數

지금 j번째 변수 X_j의 i번째 측정에서 誤差 e_{ij}가 발생한다고 하자. 또 e_{ij}에 대하여 다음

* 永薰高等學校敎師

과 같은 가정을 하자.

(i) $e_{ij}\sim N(0,\sigma_j^2)$, $\sigma_j^2<\infty, i=1,2,\cdots,q.$

(ii) $E(e_{ij}e_{ik})=0$, $E(e_{ij}e_{kj})=0$, $i\neq k$, $j\neq k$. [2]

그러면 誤差를 고려한 행렬 X를 X_e로 표시하면

$$(5)\quad \boldsymbol{X}_e=\begin{bmatrix} x_{11} & x_{12} & \cdot\cdot & x_{1q} \\ x_{21} & x_{22} & \cdot\cdot & x_{2q} \\ \cdot & \cdot & \cdot & \\ x_{n1} & x_{n2} & \cdot\cdot & x_{nq} \end{bmatrix}+\begin{bmatrix} e_{11} & e_{12} & \cdot\cdot & e_{1q} \\ e_{21} & e_{22} & \cdot\cdot & e_{2q} \\ \cdot & \cdot & \cdot & \\ e_{n1} & e_{n2} & \cdot\cdot & e_{nq} \end{bmatrix}$$

이며 간단히 표시하면 다음과 같다.

$$(6)\quad \boldsymbol{X}_e=\boldsymbol{X}+\boldsymbol{e},\ \text{단 } \boldsymbol{e}=\begin{bmatrix} e_{11} & e_{12} & \cdot\cdot & e_{1q} \\ e_{21} & e_{22} & \cdot\cdot & e_{2q} \\ \cdot & \cdot & \cdot & \\ e_{n1} & e_{n2} & \cdot\cdot & e_{nq} \end{bmatrix}$$

따라서 $\boldsymbol{X}_e$로 구한 推定量 $\hat{\boldsymbol{\beta}}(e)$는

$$(7)\quad \begin{aligned}\hat{\boldsymbol{\beta}}(e)&=(\boldsymbol{X}_e'\boldsymbol{X}_e)^{-1}\boldsymbol{X}_e'\boldsymbol{y}\\ &=[(\boldsymbol{X}'+\boldsymbol{e}')(\boldsymbol{X}+\boldsymbol{e})]^{-1}(\boldsymbol{X}'+\boldsymbol{e}')\boldsymbol{y}\\ &=[\boldsymbol{X}'\boldsymbol{X}+(\boldsymbol{e}'\boldsymbol{X}+\boldsymbol{X}'\boldsymbol{e}+\boldsymbol{e}'\boldsymbol{e})]^{-1}(\boldsymbol{X}'+\boldsymbol{e}')\boldsymbol{y}\\ &=(\boldsymbol{S}+\boldsymbol{D})^{-1}(\boldsymbol{X}'+\boldsymbol{e}')\boldsymbol{y}\\ &=(\boldsymbol{I}_q+\boldsymbol{S}^{-1}\boldsymbol{D})^{-1}\boldsymbol{S}^{-1}(\boldsymbol{X}'+\boldsymbol{e}')\boldsymbol{y}\end{aligned}$$

단 $\boldsymbol{S}=\boldsymbol{X}'\boldsymbol{X}$. $\boldsymbol{D}=\boldsymbol{e}'\boldsymbol{X}+\boldsymbol{X}'\boldsymbol{e}+\boldsymbol{e}'\boldsymbol{e}$

정의 $m\times n$ 행렬의 무한급수

$$\sum_{k=0}^{\infty}\boldsymbol{A}^{(k)}=\boldsymbol{A}^{(0)}+\boldsymbol{A}^{(1)}+\boldsymbol{A}^{(2)}+\cdots+\boldsymbol{A}^{(p)}+\cdots$$

가 행렬 $\boldsymbol{A}$에 수렴하는 필요충분 조건은 모든 $i=1,2,\ldots,n$, $j=1,2,\ldots,n$에 대하여 급수 $\sum_{k=0}^{\infty}a_{ij}^{(k)}$가 a_{ij}에 수렴하는 것이다. 단 여기서 $\boldsymbol{A}^{(k)}=(a_{ij}^{(k)})$, $\boldsymbol{A}=(a_{ij})$임을 표시한다. [3]

이제 (7)식에서 $(\boldsymbol{I}_q+\boldsymbol{S}^{-1}\boldsymbol{D})^{-1}\boldsymbol{S}^{-1}$를 전개하면

$$(8)\quad (\boldsymbol{I}_q+\boldsymbol{S}^{-1}\boldsymbol{D})^{-1}\boldsymbol{S}^{-1}=\left\{\sum_{k=0}^{\infty}(-1)^k(\boldsymbol{S}^{-1}\boldsymbol{D})^{(k)}\right\}\boldsymbol{S}^{-1}$$

($k\geqq 3$인 경우를 무시하면)

$$\fallingdotseq \boldsymbol{S}^{-1}-\boldsymbol{S}^{-1}\boldsymbol{D}\boldsymbol{S}^{-1}+\boldsymbol{S}^{-1}\boldsymbol{D}\boldsymbol{S}^{-1}\boldsymbol{D}\boldsymbol{S}^{-1}$$

(8)식을 (7)식에 대입하여 정리하고 $e^{(3)}$항 이상을 무시하여 근사값을 구하면 다음과 같다.

$$(9)\quad \begin{aligned}\hat{\boldsymbol{\beta}}(e)&\fallingdotseq(\boldsymbol{S}^{-1}-\boldsymbol{S}^{-1}\boldsymbol{D}\boldsymbol{S}^{-1}+\boldsymbol{S}^{-1}\boldsymbol{D}\boldsymbol{S}^{-1}\boldsymbol{D}\boldsymbol{S}^{-1})(\boldsymbol{X}'+\boldsymbol{e}')\boldsymbol{y}\\ &=\hat{\boldsymbol{\beta}}-\boldsymbol{S}^{-1}\boldsymbol{D}\hat{\boldsymbol{\beta}}+\boldsymbol{S}^{-1}\boldsymbol{D}\boldsymbol{S}^{-1}\boldsymbol{D}\hat{\boldsymbol{\beta}}+\boldsymbol{S}^{-1}\boldsymbol{e}'\boldsymbol{y}\end{aligned}$$

$$
\begin{aligned}
&-S^{-1}DS^{-1}e'y+S^{-1}DS^{-1}DS^{-1}e'y\\
&\fallingdotseq\hat{\beta}-S^{-1}(e'X+X'e+e'e)\hat{\beta}\\
&\quad+S^{-1}(e'X+X'e+e'e)S^{-1}(e'X+X'e+e'e)\hat{\beta}\\
&\quad+S^{-1}e'y-S^{-1}(e'X+X'e+e'e)S^{-1}e'y\\
&\fallingdotseq\hat{\beta}-S^{-1}(e'X+X'e+e'e)\hat{\beta}\\
&\quad+S^{-1}(e'X+X'e)S^{-1}(e'X+X'e)\hat{\beta}\\
&\quad+S^{-1}e'y-S^{-1}(e'X+X'e)S^{-1}ey\\
&=\hat{\beta}+S^{-1}e'(y-X\hat{\beta})-S^{-1}X'e\hat{\beta}\\
&\quad-S^{-1}e'(I_n-XS^{-1}X')e\hat{\beta}+S^{-1}X'eS^{-1}X'e\hat{\beta}\\
&\quad-S^{-1}(e'X+X'e)S^{-1}(y-X\hat{\beta})
\end{aligned}
$$

더우기 $e^{(2)}$항 이상을 무시하면 다음과 같은 근사적인 推定量을 구할 수 있다.

(10) $\hat{\beta}(e)\fallingdotseq[I_q-(X'X)^{-1}e'X-(X'X)^{-1}X'e-(X'X)^{-1}e'e]\hat{\beta}+(X'X)^{-1}e'y$ 여기서 $e^{(2)}$의 주요항 $-S^{-1}e'e\hat{\beta}$를 고려하면

(10)′ $\hat{\beta}(e)\fallingdotseq[I_q-(X'X)^{-1}e'X-(X'X)^{-1}X'e-(X'X)^{-1}e'e]\hat{\beta}+(X'X)^{-1}e'y$

정리 1 $\hat{\beta}(e)$의 근사공식으로 (10)′을 쓰면

$$E[\hat{\beta}(e)]=[I_q-n(X'X)^{-1}V]\beta,\quad V=\begin{pmatrix}\sigma_1^2 & & 0\\ & \ddots & \\ 0 & & \sigma_q^2\end{pmatrix}$$

증명 $\hat{\beta}=\beta+(X'X)^{-1}X'u$, $y-X\hat{\beta}=[I_n-X(X'X)X'^{-1}]u$를 (10)′식에 대입하여 정리하면

$$
\begin{aligned}
\hat{\beta}(e)=&[I_q-(X'X)^{-1}e'X-(X'X)^{-1}X'e-(X'X)^{-1}e'e][\beta+(X'X)^{-1}X'u]\\
&+(X'X)^{-1}e'X[\beta+(X'X)^{-1}Xu]+(X'X)^{-1}e'u\\
&-(X'X)^{-1}e'X(X'X)^{-1}X'u
\end{aligned}
$$

따라서

$E[\hat{\beta}(e)]=\beta+(X'X)^{-1}E(e'e)\beta=[I_q-n(X'X)^{-1}V]\beta$ 위에서 알 수 있는 바와 같이 $\hat{\beta}(e)$는 偏倚를 갖는 推定量이다.

정리 2 $\hat{\beta}(e)\fallingdotseq\beta+(X'X)^{-1}X'u-(X'X)^{-1}X'e\beta$를 사용하여 미소항($e_{ij}^3$, u_i^3항) 이상을 무시하면

$$V(\hat{\beta}(e))\fallingdotseq(X'X)^{-1}\left\{\sum_{i=1}^{\infty}\beta_i^2\sigma_i^2+\sigma^2\right\}$$

증명 $V(\hat{\beta}(e))=E[(\hat{\beta}(e)-\beta)(\hat{\beta}(e)-\beta)']$에서

$$
\begin{aligned}
V(\hat{\beta}(e))&\fallingdotseq(X'X)^{-1}E(uu')+(X'X)^{-1}X'E(e\beta\beta'e')X(X'X)^{-1}\\
&=(X'X)^{-1}\sigma^2+(X'X)^{-1}X'\Big(\sum_{i=1}^{q}\sigma_i^2\beta_i^2\Big)I_nX(X'X)^{-1}\\
&=(X'X)^{-1}\sigma^2+(X'X)^{-1}X'I_nX(X'X)^{-1}\Big(\sum_{i=1}^{q}\beta_i^2\sigma_i^2\Big)
\end{aligned}
$$

$$=(X'X)^{-1}\left\{\sum_{i=1}^{q}\beta_i^2\sigma_i^2+\sigma^2\right\}$$

주의 측정치에 오차가 발생하면 주어진 線型回歸模型에서 誤差項 u이외에 또 다른 誤差項 e가 생기므로 정확한 β의 推定量이 구해지지 않는다. 또 실제로 e_{ij}들을 알 수 없는 것이므로 대개 오차의 한계치를 사용하게 되며 특히 σ_j^2, $j=1,2,..,g$가 0에 대단히 가까운 값을 가지면 誤差를 고려하지 않는 경우와 거의 일치하게 되므로 자료의 정확한 측정이 중요하다.

參 考 文 獻

[1] Taylor, L, D., *Probability and Mathematical Statisitics,* New York:Harper & Row,1974

[2] Draper, N. R. and Smith H., *Applied Regession Analysis,* New York: John Wiley, 1966

[3] Finkbeiner, D.T., *Introduction to Matrices and Linear Transformations,* New York: W. H. Freeman Co. 1963

[4] Graybill, F.A., *An Introduction to Linear Statistical Models,* Vol. 1, New York: Mcgraw-Hill Inc. 1961

〈**ABSRACT**〉

Data Errors and Regression Analysis

S. K. Kim

This paper considers the problem of estimating $\hat{\beta}$ in the case errors occur in observing the values of q-variables $X_1, X_2, \cdots, X_q$. The approximated estimator $\hat{\beta}(e)$ is obtained and its expected value, bias and covariance matrix are studied.