韓國數學敎育學會誌 시리즈 A 〈數學敎育〉
1993. 12. 제 32권 제 2호 179-189

Journal of the Korea Society of Mathematical Education
Series A : Dec. 1993, Vol. 32. No. 2, 179-189

誤差項이 回歸係數에 미치는 영향

崔昌浩(江南大學校)

1. 序論

선형회귀모형

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + \varepsilon_i \qquad \cdots (1.1)$$

에서 오차항 ε_i, $(i=1, 2, \cdots, n)$ 가 선형회귀모형의 기본가정

$$\varepsilon_i \sim i.i.d.\ N(0, \sigma^2),\ (i.\ i.\ d.\ :\ \text{independently identical distribution})$$

을 만족하였을 때 최소자승법(O.L.S. : ordinary least square)으로 추정한 회귀계수의 추정량 $\hat{\beta}_i, (i=0,1,2,\cdots,k)$는 최소분산 일치추정량이 되었다.

그러나 실제로는 이와같은 가정을 만족하지 않는 경우가 가끔 나타나며 이 경우

$$\underline{\hat{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_k)'$$

단 : $= (X'X)^{-1}X'\underline{y}$

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

$$X = \begin{bmatrix} 1 & x_{11} & x_{21} & \cdots & x_{k1} \\ 1 & x_{12} & x_{22} & \cdots & x_{k2} \\ & \cdots & & \cdots & \\ 1 & x_{1n} & x_{2n} & \cdots & x_{kn} \end{bmatrix}$$

는 최소분산 일치추정량이 되지 못한다. 예를들면 오차항이

$$\varepsilon_i = \rho\varepsilon_{i-1} + \delta_i, \qquad 단 : \ \delta_i \sim i.i.d.\ N(0,\ \sigma_\delta^2)$$

와 같은 일계의 자기상관이 있을 가능성이 있을 때 이의 검정을 위해서는 Durbin - Watson통계량이 이용되며 검정의 결과 일계의 자기상관이 있음이 확인되면 회귀계수의 추정을 위하여는 적당한 변환이 필요하게 된다.

본 논문에서는 기본가정

$$\varepsilon_i \sim i.i.d.\ N(0,\ \sigma^2)$$

이 만족되지 않는 경우 회귀계수의 일치추정량을 구하는 방법에 대해 알아보고자 한다. 아울러 어느 설명변수가 반응변수에 더많은 영향을 미치는가를 알아 볼 수 있는 방법을 제시하고자 한다.

2. 時系列 模形

시계열 Z_t가 정상(stationary)이거나 비정상(nonstationary) ARIMA(auto-regressive integrated moving average model)(p. d. q.)모형을 따르면

$$\Phi(B)Z_t = \theta(B)a_t \qquad \cdots (2.1)$$

로 표시되며 여기서

$$\Phi(B) = (1-B)^d\phi(B)$$
$$\phi(B) = 1-\phi_1 B-\phi_2 B^2 -\cdots-\phi_p B^p$$
$$\theta(B) = 1-\theta_1 B-\theta_2 B^2 -\cdots-\theta_q B^q$$
$$B^m Z_t = Z_{t-m},\ (m=0,\ 1,\ 2,\ \cdots)$$

이고 $\{a_t\}$는 백색잡음과정(white noise process)로 $E(a_t)=0,\ \mathrm{var}(a_t)=\sigma_a^2$이다.

그리고 (2. 1)은 다음 조건을 만족한다.

$\phi(B)=0$ 의 근은 단위원 밖에 존재하고 $\theta(B)=0$의 근은 단위원 위이거나 밖에 존재하며 $\phi(B)=0$ 과 $\theta(B)=0$은 공통근을 갖지 않는다.

ARIMA(p. d. q.)모형에서

$d=0$이면 정상시계열모형(stationary time series model),

$d\geq 1$인 정수이면 비정상시계열모형(nonstationary time series model),

$d=0, q=0$ 이면 차수가 p인 자기회귀모형(pth autoregressive model),

$d=0$, $p=0$ 이면 차수가 q인 이동평균모형(qth moving average model)

이라고 한다.

다음 몇가지 정의를 밝히고자 한다.

〔정의 2. 1〕

시차 k인 표본자기상관함수(SACF : sample autocorrelation funtion)는

$$r(k)=\frac{\sum_{t=k+1}^{n}(Z_t-\overline{Z})(Z_{t-k}-\overline{Z})}{\sum_{t=1}^{n}(Z_t-\overline{Z})^2},\quad k=0,\ 1,\ 2,\ \cdots$$

단: $\overline{Z}=\sum_{t=1}^{n}Z_t/n$

로 정의 한다.

〔정의 2.2〕

차수가 k인 자기회귀모형

$$Z_t=\phi_{k1}Z_{t-1}+\cdots+\phi_{kk}Z_{t-k}+a_t$$

로 부터 얻은 Yule - Walker 방정식

$$\rho_1 = \phi_{k1} + \phi_{k2}\rho_1 + \cdots + \phi_{kk}\rho_{k-1}$$
$$\rho_2 = \phi_{k1}\rho_1 + \phi_{k2} + \cdots + \phi_{kk}\rho_{k-2}$$
$$\cdots \qquad\qquad \cdots$$
$$\rho_k = \phi_{k1}\rho_{k-1} + \phi_{k2}\rho_{k-2} + \cdots + \phi_{kk}$$

의 ρ_k 대신 표본자기상관함수값 r_k를 대입한 연립방정식으로 부터 구한 $\phi_{kk},(k=0,1,2,\cdots)$를 표본편자기상관함수(SPACF : sample partial autocorrelation funtion)이라고 한다.

〔정의 2.3〕

$$\hat{\phi}_i^{(j)},\ i=1,\ 2,\ \cdots,\ m,\ \ m=1,\ 2,\ \cdots$$

를 차수가 m인 자기회귀모형

$$Z_t = \phi_1 Z_{t-1} + \phi_2 Z_{t-2} + \cdots + \phi_m Z_{t-m} + a_t$$

단 : a_t는 오차항

의 j번째 반복회귀(iterated regression)

$$Z_t = \sum_{j=1}^{m} \phi_i^{(j)} Z_{t-i} + \sum_{i=1}^{j} \beta_i^{(j)} \hat{a}_{t-i}^{(j-i)} + a_t^{(j)}$$
$$t = m+j+1,\ \ m+j+2,\ \cdots,\ n$$

에 최소자승법을 이용하여 얻은 최소자승추정값이라고 할 때 Z_t의 m번째 확장된 표본자기상관함수(ESACF : Extended sample autocorrelation funtion) $\hat{\rho}_j^{(m)}$ 는 변환된 시계열

$$\hat{Y}_t^{(j)} = (1 - \hat{\phi}_1^{(j)}B - \hat{\phi}_2^{(j)}B^2 - \cdots - \hat{\phi}_m^{(j)}B^m)Z_t$$

그리고 차수 $p,\ d,\ q.$ 의 결정은 먼저

$$(1-B)^d Z_t = w_t$$

가 정상시계열이 되도록 차분의 차수 d를 정한다음 표본자기상관함수와 표본편자기 함수 및 확장된 표본자기상관함수를 이용하여 자기회귀부분의 차수 p와 이동평균부분 의 차수 q를 결정한다.

〔정의 2.4〕
주어진 $i(1 \le i \le k)$ 에 대하여

$$\delta_{ii} = \min_{0 \le \delta < \infty} \left\{ \delta \Big|^{\min}_{n \to \infty} n^{-\delta} \sum_{t=1}^{n} X_{it}^2 < \infty \right\} \qquad \cdots (2.2)$$

인 음이 아닌 실수 δ_{ii} 가 존재하며 이를 변수 X_{it}의 차수(order)라고 한다.
(2.1)식, 즉 ARIMA(p, d, q.)모형

$$(1-B)^d \Phi(B) Z_t = \theta(B) a_t$$

에서는

$$\left.\begin{aligned} &\sum_{t=1}^{n} Z_t^2 = O_p(n^{2d}) \\ &\left[\sum_{t=1}^{n} Z_t^2\right]^{-1} = O_p(n^{-2d}) \\ &\sum_{t=1}^{n} Z_t W_{t+h} = O_p(n^{d+d'}) \end{aligned}\right] \qquad \cdots (2.3)$$

임이 Tiao와 Tsay(1983)에 의하여 밝혀졌다. 여기서 h는 주어진 상수이고

$$W_t = (1-B)^{d^*} Z_t$$

이며

$$d' = d - d^*$$

이다.

그리고 Z_t의 표본자기상관함수 $r_Z(l)$는 동차차분방정식(homogeneous difference equation)

$$(1-B)r_Z(l)=0,\ (l\text{: 고정된 정수})$$

를 만족한다.

또 (2.3)식으로 부터 $d \geq 1$이면

$$\left.\begin{aligned} &i)\ \sum_{t=1}^{n}(Z_t-\bar{Z})^2=O_p(n^{2d}) \\ &ii)\ \left[\sum_{t=1}^{n}(Z_t-\bar{Z})^2\right]^{-1}=O_p(n^{-2d}) \\ &iii)\ \sum_{t=1}^{n}(Z_t-\bar{Z})(W_{t+h}-\bar{W})=O_p(n^{d+d'}) \end{aligned}\right] \quad \cdots (2.4)$$

단 : $\bar{Z}=\sum_{t=1}^{n}Z_t/n\ ,\qquad \bar{W}=\sum_{t=1}^{n}W_t/n$

가 성립함을 쉽게 알 수 있다.

3. 定理

선형회귀모형의 오차항이 ARIMA($p.\ d.\ q.$)모형

$$\Phi(B)\varepsilon_t=\theta(B)a_t$$

를 따른다면 (1.1)식은

$$Y_t=\sum_{i=1}^{k}\beta_i X_{it}+[\Phi(B)]^{-1}\theta(B)a_t \quad \cdots (3.1)$$

와 같이 표시된다.

그리고 (2.2)에서 정의된 $\delta_{ii}(i=1, 2, \cdots, k)$가운데 서로 다른 값들이 r개가 있다면 다음과 같이 크기의 순서로 배열하도록 한다.

$$\delta_1 \rangle \delta_2 \rangle \cdots \rangle \delta_r$$

그러면 다음과 같은 정리들이 성립한다.

〔정리 3.1〕

Y_t가 (3.1)식과 같은 선형회귀모형을 따르고 δ_1을 $X_{it}(i=1, 2, \cdots, k)$들의 차수가운데 최고위차수라고 할 때 $2d \rangle \delta_1$이면 표본자기상관함수(SACF) $r_y(l)$는 근사적으로 동차차분방정식

$$(1-B)r_y^{(l)} = 0, \quad (l: \text{임의의 고정된 정수})$$

을 만족한다.

〔증명〕

(3.1)식으로 부터

$$\overline{Y} = \sum_{i=1}^{k} \beta_i \overline{X}_i + \overline{\varepsilon}$$

단 : $\overline{Y} = \sum_{t=1}^{n} Y_t /n, \quad \overline{X}_i = \sum_{t=1}^{n} X_{it} /n, \quad \overline{\varepsilon} = \sum_{t=1}^{n} \varepsilon_t / n$

이므로

$$Y_t - \overline{Y} = \sum_{i=1}^{k} \beta_i (X_{it} - \overline{X}_i) + (\varepsilon_t - \overline{\varepsilon})$$

이다. 따라서 (2.4)식에 의하여

$$\sum_{t=1}^{n}(Y_t-\overline{Y})^2=O_p(n^{2d})$$

$$\left[\sum_{t=1}^{n}(Y_t-\overline{Y})^2\right]^{-1}=O_p(n^{-2d})$$

가 성립된다. 또

$$V_t=(1-B)Y_t\,,\quad W_t=(1-B)\varepsilon_t$$

로 놓으면

$$V_t=\sum_{i=1}^{k}\beta_i\left[(1-B)X_{it}\right]+W_t$$

이므로

$$(1-B)r_y(l)\doteq\left[\sum_{t=1}^{n}(Y_t-\overline{Y})\right]^{-1}\left[\sum_{t=l+2}^{n}(Y_t-\overline{Y})(V_t-\overline{V})\right]$$

이다. 그리고 $2d\rangle\delta_1$ 이므로

$$\sum_{t=l+2}^{n}(V_t-\overline{V})^2=O_p(n^{\delta})$$

단 : $\delta=\max\{\delta_1,\ 2(d-1),\ 1\}\langle 2d$

이다. 이상으로 부터

$$(1-B)r_y(l)=O_p(n^{-d+\frac{\delta}{2}})$$

이므로 본 정리는 성립된다.

〔정리 3.2〕

시계열 Y_t가 (3.1)식과 같은 선형회귀모형을 따르고 적당한 j에 대하여 $2d < \delta_j \ (1 \le j \le k)$이라 하자.

그리고 δ_{ii} 를 설명변수 X_{it}의 차수라고 할 때

i) $\delta_{ii} > 2d$ 이면 설명변수 X_{it}의 회귀계수 β_i 최소자승추정값 $\hat{\beta}_i$는 β_i의 일치추정값이 되고

ii) $\delta_{ii} < 2d$이면 최소자승추정값 $\hat{\beta}_i$는 β_i의 일치추정값이 아니며

iii) $\delta_{ii} = 2d$이면 최소자승추정값 $\hat{\beta}_i$가 β_i의 일치추정값인지의 여부를 판정할 수 없다.

〔증명〕

증명을 간단히 하기 위하여 k개의 설명변수가운데 b개의 차수는 δ_1 이고 나머지 $c = k - b$개의 차수는 δ_2 라고 하면

$$\delta_1 > 2d \ge \delta_2$$

이다. 그러면 최소자승추정값의 추정오차는

$$\underline{\hat{\beta}} - \underline{\beta} = (\sum_{t=1}^{n} X_t X_t')^{-1} (\sum_{t=1}^{n} X_t \varepsilon_t)$$

이며 II_j를 $j \times j$인 단위행렬. $G_1 = n^{\delta_1} II_b$, $G_2 = n^{\delta_2} II_c$ 이라하고 $G = diag\{G_1, G_2\}$이라면

$$\underline{\hat{\beta}} - \underline{\beta} = (G_1^{-1} \sum_{t=1}^{n} X_t X_t')^{-1} (G^{-1} \sum_{t=1}^{n} X_t \varepsilon_t)$$

$$= \begin{bmatrix} O(1) & O(n^{-\omega}) \\ O(n^{\omega}) & O(1) \end{bmatrix}^{-1} \begin{bmatrix} O_p(n^{-\lambda}) \\ O_p(n^{-\gamma}) \end{bmatrix}$$

단 : $\omega = (\delta_1 - \delta_2)/2$, $\lambda = (\delta_1 - 2d)/2$, $\gamma = (\delta_2 - 2d)/2$

이며 행렬의 분할은 차수 b와 c에 의하여 이루어졌다.
이다. 그리고

$$-\omega-\gamma=-(\delta_1-2d)/2\langle 0 \quad \text{이고} \quad -\lambda\langle 0$$

이므로 정리의 내용가운데 i)이 설명변수의 차수가 δ_1 과 δ_2 두개뿐인 경우에 대해서 성립되며 이를 일반화함으로서 i)이 성립하게 된다.

다음으로 ii)와 iii)은

$$\lambda=-\gamma=2d-\delta_2 \geq 0$$

로 부터 성립됨을 쉽게 알 수 있다.

참고문헌

Anderson, O. D., Anaysing time series, North Holland publishing company, 1980.

Box, G. E. P., and Jenkins, G. M., Time Series Analysis-Forecasting and Control, Holden-Day, 1976.

Fuller W. A., Introduction to statistical time series, John Wiley and Sons, Inc., 1976.

Montgomery D. C. and Johnson L. A., Forecasting and time series analysis, McGraw-Hill, 1976.

Nelson C. R., Applied time series analysis, Holsen-Day, 1973.

Ruey S. T,. "Regression Models With Time Series Errors, Journal of the American Statistical Association, Volume 79, No. 385, 1984.

Tsay, R. S., and Tiao, G. C., "Consistent Estimates of Autoregressive Parameters and Extended Sample Autocorrelation Function for Stationary and Nonstationary ARMA Models, Journal of the American Statistical Association, 79, 84-96, 1984.