응용통계연구 제 **16**권 2호, *2003*년, *pp. 365-376*

비정규 시계열 자료의 회귀모형 연구 *

최윤하[1] 이성임[2] 이상열[3]

요 약

본 연구에서는 비정규 시계열 자료에 관한 다양한 회귀모형을 고찰하고, 이들 모형의 선택 기준에 관하여 연구해 보았다. 모형 선택의 기준으로는 AIC (Akaike information criterion), BIC (Baysian information criterion) 그리고 우도비 검정을 확장 적용하였다. 또한, 실제의 Polio 자료분석을 통해 이를 적용해보았다.

주요용어: 비정규 시계열 자료 , 회귀모형, 모형선택

1. 서론

지금까지 정규성을 가정할 수 있는 시계열 자료에 관한 회귀모형에 비해, 정규성을 만족하지 않는 예를 들어, 지수분포족을 따르는 시계열 자료의 회귀모형은 상대적으로 많은 연구가 이루어지지 않았다. 그러나, 실질적으로는 "예/아니오" 의 범주형으로 기록된 일별 강우량이나 임의의 질환에 대한 환자의 월별 발생수 등과 같이 정규성을 가정하기 어려운 시계열 자료가 많이 있고, 통계학을 응용하는 분야가 넓어지면서 이러한 자료를 다루는 회귀모형의 필요성이 날로 커지고 있다.

고전적인 회귀모형과 마찬가지로 시계열 자료 $(y_t, \mathbf{x}_t)$에 기초한 회귀 모형은 y_t에 미치는 $\mathbf{x}_t$의 영향이 주된 관심이다. 그런데, 기존의 회귀모형은 자료의 독립성을 가정하고 있는 반면, 시계열 자료의 경우에는 계열상관 (serial correlation)이 존재하기 마련이다. 그러므로, 시계열 자료의 회귀모형에서는 관측값들의 자기상관을 설명하는 것이 선결과제이다. 이를 해결하는 방법으로는 크게 두 가지를 생각해 볼 수 있다. 첫째는 몇 개의 과거 관측값을 예측변수 (predictor variable)로 취함으로써 시계열 자료가 갖는 자기상관을 설명하려는 것이다. 즉, 과거의 관측값 $y_{t-1}, \ldots, y_1$과 공변량 $\mathbf{x}_t$가 주어졌을 때, 이들에 대한 y_t의 조건부 평균 (τ_t)을 모형화 하는 것이다. Brumback et al.(2000)은 이러한 접근법을 전이 회귀 함수 (transitional regression model)라는 모형으로 설명하였다. 이와 관련한 대표적 모형으로는 Cox (1970), Korn & Whittemore (1979), 그리고 Kalbfleisch & Lawless (1985)와 거기에 나와있는 참고문헌을 참조하기 바란다. 또 다른 접근법은 잠재적인 자기상관 확률과정 ϵ_t

* 본 연구는 서울대학교 복잡계통계연구센터를 통한 한국과학재단의 지원에 의하여 수행되었음.

1) (151-742) 서울시 관악구 신림동 산56-1, 서울대학교 통계학과, 석사
E-mail : yhchoi@stats.snu.ac.kr
2) (151-742) 경기도 안산시 단원구 고잔 1동 516, 고려대학교 의과학연구원, 연구조교수
E-mail : silee70@kumc.ns.or.kr
3) (151-742) 서울시 관악구 신림동 산56-1, 서울대학교 통계학과, 교수
E-mail : sylee@stats.snu.ac.kr

를 갖는 구조적 모형 (structural model)을 세우는 것이다. 즉, ϵ_t가 주어졌을 때 y_t는 서로 독립이라 가정할 수 있고, 따라서 $\mathbf{x}_t$만 가지고 y_t의 조건부 평균 (μ_t) 을 모형화 하는 것이다. 이와 관련하여 Samet et al. (1995)과 Zeger (1988)등을 참조하기 바란다. Cox (1981)는 첫 번째 방법을 "관측치에 의한 모형 (Observation-Driven Model; ODM)으로 두 번째 방법을 "모수에 의한 모형 (Parameter-Driven Model; PDM)이라 소개하였다.

본 연구에서는 지금까지 연구된 비정규성 시계열 자료에 기초한 다양한 회귀 모형들에 대하여 고찰하고, 모형 선택 기준으로 대표적인 검정 통계량들을 비정규 시계열 자료의 모형으로 확장 적용할 것이다. 또한, 실제의 자료 분석을 통해, 주어진 자료로부터 지금까지 알려진 적합 가능한 모든 모형을 제시하고, 이들로부터 어떤 모형이 최종 선택되는지 분석해 보았다.

2. 비정규 시계열자료를 위한 회귀 모형

시계열 자료는 자료의 독립성을 가정할 수 없기 때문에, 독립성을 기본으로 하는 일반적인 회귀 모형을 그대로 적용할 수가 없다. 이 절에서는 이러한 비독립적인 자료에 기초하여, 정규성을 따르지 않는 경우의 회귀 분석을 위해 지금까지 제안된 모형들을 알아보기로 한다. 비정규 시계열자료를 위한 회귀모형은 크게 조건부 모형(Conditional Model)과 주변 모형(Marginal Model)로 분류할 수 있다. 조건부 모형은 적절한 시차 (lag)의 과거 관측값을 공변량 $\mathbf{x}_t$와 함께 설명변수로 사용하여 자료의 자기상관을 설명한 가법 모형이다. 한편, 주변 모형은 관측할 수 없는 잠재된 확률과정 ϵ_t를 곱의 형태로 적용하여, ϵ_t를 통해 자료의 자기상관을 설명한 승법 모형이라 할 수 있다. 다음에서 각 모형에 대해 자세히 알아보기로 하자.

2.1. 조건부 모형(Conditional Model)

조건부 모형이란 자료의 자기상관을 고려하기 위하여, 자료를 과거의 관측값에 의존하여 설명한 모형이다. 이 모형에서는 기본 가정으로 y_t의 평균과 분산이 과거의 관측값 $y_{t-1},\ldots,y_1$과 공변량 $\mathbf{x}_t$에 대한 조건부 형태로 주어진다. $\mathbf{H}_t$를 $\{y_{t-1},\ldots,y_1,x_1,\ldots,x_{t-1}\}$의 함수 형태로 주어질 때, 이를 과거함수(history function)라 하고, 조건부 모형의 기본 정의를 살펴보면 다음과 같다.

- 모형 1

$$\begin{aligned} y_t &= \tau_t + \delta_t, \qquad t = 1,\cdots,n, \\ g(\tau_t) &= \mathbf{x}_t\beta + \mathbf{H}_t\alpha, \\ \tau_t &= E(y_t|\mathbf{x}_t,\mathbf{H}_t), \end{aligned} \tag{2.1}$$

여기서 y_t의 $\mathbf{x}_t,\mathbf{H}_t$에 대한 조건부 분산은 $a(\tau_t)$가 된다. 따라서 $var(\delta_t|\mathbf{x}_t,\mathbf{H}_t) = a(\tau_t)$가 된다. 만약 $\mathbf{H}_t$가 $\{y_{t-1},\cdots,y_{t-p},\mathbf{x}_{t-1},\cdots,\mathbf{x}_{t-p}\}$로 이루어져 있다면, 이는 시차가 p인 모형이라고 한다. 일반적으로 $\mathbf{H}_t = (H_{t-1},\cdots,H_{t-p})$와 같이 구성되며 H_{t-p}는 y_{t-p}와 $\mathbf{x}_{t-p}$로 표

현되는 과거함수 (history function)이다. 이 모형은 전이 일반화 선형모형 (the transitional generalized linear model ;TGLM)(Brumback, 2000)이라 부르기도 한다.

그리고 $\mathbf{H}_t$의 형태에 따라 평균형과 오잔차형 함수로 분류할 수 있는데, 그 형태는 다음과 같다. 모든 $\mathbf{H}_t$의 시차는 p로 표현하였다.

표 1. 과거함수의 형태 분류

$\mathbf{H}_t$	평균형	오차형
Type 1	$(\ y_{t-1} \ , \ \cdots \ , \ y_{t-p} \ )$	$(\ y_{t-1} - \tau_{t-1} \ , \ \cdots \ , \ y_{t-p} - \tau_{t-p} \ )$
Type 2	$(\ \tau_{t-1} \ , \ \cdots \ , \ \tau_{t-p} \ )$	$(y_{t-1} - g^{-1}(\mathbf{x}_{t-1}\beta), \cdots, y_{t-p} - g^{-1}(\mathbf{x}_{t-p}\beta))$
Type 3	$(g^{-1}(\mathbf{x}_{t-1}\beta), \cdots, g^{-1}(\mathbf{x}_{t-p}\beta))$	$(g(y_{t-1}) - \mathbf{x}_{t-1}\beta \ , \ \cdots \ , \ g(y_{t-p}) - \mathbf{x}_{t-p}\beta)$
Type 4		$(\ g(y_{t-1}) - \tau_{t-1} \ , \ \cdots \ , \ g(y_{t-p}) - \tau_{t-p} \ )$

평균형의 Type 1은 관측값 $(y_{t-1}, \cdots, y_{t-p})$을 그대로 과거 함수 $\mathbf{H}_t$로 사용하는 것이다. Type 2는 $(\tau_{t-1}, \cdots, \tau_{t-p})$, 즉 $(g^{-1}(\mathbf{x}_{t-1}\beta + \mathbf{H}_{t-1}\alpha), \cdots, g^{-1}(\mathbf{x}_{t-p}\beta + \mathbf{H}_{t-p}\alpha))$를 $\mathbf{H}_t$로 하는 것이다. 이는 Type 3인 $(g^{-1}(\mathbf{x}_{t-1}\beta), \cdots, g^{-1}(\mathbf{x}_{t-p}\beta))$와 유사한 형태이나 Type 3과는 달리 연결함수의 역함수 g^{-1}안에 $\mathbf{H}_{t-i}\alpha$항이 포함되어 있다는 차이가 있다. 오차형의 Type 1은 실제 오차값을 과거함수로 사용한 것이며, 오차형 Type 2는 평균형의 Type 3 처럼 τ_{t-i}대신에 $g^{-1}(\mathbf{x}_{t-i}\beta)$를 사용한 것이다.

특히, 표1 에서 평균형의 Type 1, $\mathbf{H}_t = (y_{t-1}, \cdots, y_{t-p})$일 경우, 그 모형을 일반화 자기회귀 모형(generalized autoregressive models)이라 한다. 또한, 잔차형의 Type 3과 4를 사용한 경우의 모형은 일반화 자기회귀 평균이동 모형 (generalized autoregressive moving average models:GARMA)라 부른다. GARMA모형은 다음과 같다.:

$$g(\tau_t) = \mathbf{x}_t\beta + \sum_{i=1}^{p} \alpha_i(g(y_{t-i}) - \mathbf{x}_{t-i}\beta) + \sum_{j=1}^{q} \alpha_j^{'}(g(y_{t-j}) - \tau_{t-j})$$

또한, 우리가 고려해 볼 수 있는 모형으로 과거 함수 $\mathbf{H}_t$를 오차항에 적용시키는 모형이다. 가장 일반적인 형태를 살펴보면 다음과 같다.

- 모형 2

$$\begin{aligned} y_t &= \mu_t + e_t, \qquad t = 1, \cdots, n, \\ g(\mu_t) &= g(E(y_t|\mathbf{x}_t)) = \mathbf{x}_t\beta, \\ e_t &= \alpha_1(v_t/v_{t-1})e_{t-1} + \cdots + \alpha_q(v_t/v_{t-q})e_{t-q} + \delta_t. \end{aligned} \tag{2.2}$$

이 때 μ_t는 y_t의 $\mathbf{x}_t$에 대한 조건부 평균이며 e_t는 오차항으로 평균이 0이고, 분산은 v_t이다. 따라서 $var(\delta_t) = a(\mu_t)(1-k(\alpha))$이 성립함을 쉽게 알 수 있다. 이 때, $k(\alpha)$는 $var(\alpha_1(e_{t-1}/v_{t-1}) + \cdots + \alpha_q(e_{t-q}/v_{t-q}))$이다.

모형 2는 오차 e_t가 시계열 AR(p)를 따르므로 시계열 오차를 이용한 조건부 모형이라 부른다. 또한, Brumback (2000)은 이를 시계열 오차를 이용한 일반화 선형모형 (generalized

linear model with time series errors ; GLM with TSE)이라 칭하고, 이를 위에서 소개한 모형인 전이 일반화 선형모형 (TGLM;(2.1))과 같이 전이 회귀 모형 (transitional regression model;TRM)으로 통합하여 설명하였다.

모형 2의 오차항의 부분을 오차형 과거함수(error-type history function)로 간주하면 조건부 모형 (2.1)과 같이 $(\mathbf{x}_t, \mathbf{H}_t')$에 대한 조건부 평균으로 표현할 수 있다. 즉, $\mathbf{H}_t' = ((v_t/v_{t-1})e_{t-1}, \cdots, (v_t/v_{t-q})e_{t-q})$ 라 놓으면, 모형 2는 다음과 같이 표현된다.

$$\begin{aligned} g(\mu_t) &= g(E(y_t|\mathbf{x}_t)) = \mathbf{X}_t\beta, \\ E(y_t|\mathbf{x}_t, \mathbf{H}_t') &= g^{-1}(\mathbf{X}_t\beta) + \mathbf{H}_t'\alpha. \end{aligned} \tag{2.3}$$

위에서 소개한 조건부 모형의 모수 $\alpha, '\beta$추정은 $\sum \delta_t^2$를 최소로 하는 반복재가중최소제곱법 (iteratively reweighted least squares)을 사용하였다. 본 연구에서는 기존의 S-plus의 glm()함수를 이용하여 반복추정하였다. 추정에 관련된 자세한 내용은 Fahrmeir and Tutz,G. (2001)를 참조하기 바란다.

2.2. 주변 모형(Marginal Model)

주변 모형에서는 자료의 자기상관성이 관측되지 않는 잠재된 확률과정을 통해 표현된다. g를 일반화 선형 모형에서의 연결함수라 할 때, $g(\mu_t)$은 관측되지 않은 ϵ_t에 의존하는 모수로 가정한다. 주변 모형의 예로서 Zeger (1988)가 소개한 포아송 분포를 따르는 시계열 자료에 대한 모형을 살펴보기로 하자.

시계열 y_t가 잠재된 확률 과정 ϵ_t에 대해서 조건부 독립이라고 하면 모형의 기본 구조는 다음과 같다.

- 모형 3

$$\begin{aligned} E(y_t|\epsilon_t) = var(y_t|\epsilon_t) &= \exp(\mathbf{X}_t\beta)\epsilon_t, \\ \mu_t = E(y_t) = \exp(\mathbf{X}_t\beta), \qquad & var(y_t) = \mu_t + \sigma^2\mu_t^2, \\ \rho_y(t,r) = corr(y_t, y_{t+r}) &= \frac{\rho_\epsilon(r)}{[\{1+(\sigma^2\mu_t)^{-1}\}\{1+(\sigma^2\mu_{t+r})^{-1}\}]^{\frac{1}{2}}}. \end{aligned} \tag{2.4}$$

여기서 ϵ_t는 평균이 1 이고, $cov(\epsilon_t, \epsilon_{t+r}) = \sigma^2\rho_\epsilon(r)$인 정상 확률과정이다. 또한 $\rho_\epsilon(r) = \rho^r$이라 하면, $cov(y_t, y_{t+r}) = c(\mu_t, \mu_{t+r}; \alpha)$에 대한 모수는 $\alpha = (\sigma^2, \rho)$가 된다. 모형 3의 모수 추정에 대한 자세한 방법은 Zeager (1988)을 참조하기 바란다.

3. 모형의 선택

2절에서 소개된 바와 같이 비정규 시계열 자료를 위한 회귀 모형은 자료의 자기상관성을 설명하는 방법에 따라 다양하게 존재한다. 따라서 여러 가지 가능한 모형들 중에서 자료에 가장 적합한 모형을 찾는 것은 중요한 문제이다. 이 절에서는 모형을 선택하는 방법을 제시하고자 한다. 시계열 모형에서 일반적으로 사용했던 AIC와 BIC를 확장하고, 일반

화 선형모형에서 사용되어온 우도비 검정을 확장하여 비정규 시계열 자료를 위한 회귀모형에 적용할 수 있도록 하겠다.

AIC와 BIC를 조건부 우도함수를 이용하여 모형 1에 확장 적용하면,

$$AIC = -2\sum \log L(\hat{\tau}_t, y_t|\mathbf{x}_t, \mathbf{H}_t) + 2(p_\beta + p_\alpha) \tag{3.1}$$

$$BIC = -2\sum \log L(\hat{\tau}_t, y_t|\mathbf{x}_t, \mathbf{H}_t) + log(n)(p_\beta + p_\alpha) \tag{3.2}$$

와 같이 된다. 여기서 τ_t, $\mathbf{H}_t$ 는 2절의 모형에서 설명한 바와 같이 각각 조건부 평균과 $\{y_{t-1}, \cdots, y_{t-p}, \mathbf{x}_t, \cdots, x_{t-p}\}$의 함수를 의미한다. 또한, p_β는 공변량에 대한 계수 β의 개수이며, p_α는 함수H_t에 대한 계수 α의 개수를 의미한다.

첫번째 항은 모형적합 정도의 척도 이며 두번째 항은 계수의 수에 대한 조절항으로, 모형 적합이 잘 될수록 AIC와 BIC값이 작아지게 된다. 두번째 항인 계수의 수에 대한 조절항에 있어서 AIC는 $2(p_\beta + p_\alpha)$로 사용하는 반면 BIC는 $log(n)(p_\beta + p_\alpha)$을 사용하고 있다. 따라서 BIC가 AIC보다 계수에 수에 따른 Penalty 항을 더 크게 잡고 있기 때문에 일반적으로 AIC는 변수의 개수를 상대적으로 많이, BIC는 반대로 적게 선택하는 경향이 있다.

포아송 분포를 따르는 시계열 자료에 대한 여러 회귀 모형들의 비교를 예를 들어보자. 다음의 AIC_{Diff} 와 BIC_{Diff}는 AIC와 BIC에서 모든 모형에 대해 변하지 않는 동일한 값을 갖는 항을 제하고 서로 차이를 보이는 나머지 항으로 구성한 값이다. 따라서 AIC와 AIC_{Diff} 그리고 BIC와 BIC_{Diff}는 증감 추세가 동일하게 된다. 조건부 모형인 모형 1의 경우에는

$$\begin{aligned} AIC_{Diff} &= -2\sum (y_t \log(\exp(\mathbf{x}_t\beta + \mathbf{H}_t\alpha)) - \exp(\mathbf{x}_t\beta + \mathbf{H}_t\alpha)) + 2(p_\beta + p_\alpha) \\ BIC_{Diff} &= -2\sum (y_t \log(\exp(\mathbf{x}_t\beta + \mathbf{H}_t\alpha)) - \exp(\mathbf{x}_t\beta + \mathbf{H}_t\alpha)) + \log(n)(p_\beta + p_\alpha) \end{aligned}$$

값을 계산하여 그 값을 비교하면 된다. 또한 모형 2의 경우에는 $\mathbf{H}_t$대신에 오차항에서 이끌어 낸 $\mathbf{H}_t'$를 사용한 (2.2)에서 AIC_{Diff}와 BIC_{Diff}를 계산하면 된다.

또한 주변 모형 3은 일반적인 주변 우도 함수를 사용한 AIC와 BIC를 적용한다.

$$\begin{aligned} AIC_{Diff} &= -2\sum (y_t \log(\exp(\mathbf{x}_t\beta)) - \exp(\mathbf{x}_t\beta)) + 2p \\ BIC_{Diff} &= -2\sum (y_t \log(\exp(\mathbf{x}_t\beta)) - \exp(\mathbf{x}_t\beta)) + \log(n)p \end{aligned}$$

여기서 p는 위에서와 마찬가지로 모수의 개수를 의미한다.

이러한 AIC와 BIC를 이용한 모형 비교는 서로 다른 모형간의 비교뿐만 아니라 같은 모형에서 차수를 달리한 계층적 모형들간의 비교도 가능하다. 특히 조건부 모형의 계층적 모형 선택에서는 AIC와 BIC이외에도 조건부 편차의 차를 이용할 수 있겠다. 편차의 차를 이용한 방법은 우도비 검정과 동일하다고 볼 수 있다.

계층적 모형 비교에서, 두 모형의 관계는 한 모형이 다른 모형에 포함되는 관계에 있어야 한다. 즉, 모형 A는 차수가 p-1인 관측치에 의한 모형이고, B는 차수가 p인 경우를 예로 들 수 있겠다. 포아송 자료에 대해,이 두 모형이 조건부 모형인 모형 1일 경우 편차의 차는

$$D(y; \hat{\tau}_{At}|\mathbf{x}_t, (H_{t-1}, \cdots, H_{t-p+1})) - D(y; \hat{\tau}_{Bt}|\mathbf{x}_t, (H_{t-1}, ..., H_{t-p}))$$

$$= 2\sum\left(y_t \log \frac{y_t}{\exp(\mathbf{x}_t\beta + \mathbf{H}_t(p)\alpha(p))} - (y_t - \exp(\mathbf{x}_t\beta + \mathbf{H}_t(p)\alpha(p)))\right)$$
$$- 2\sum\left(y_t \log \frac{y_t}{\exp(\mathbf{X}_t\beta + \mathbf{H}_t(p-1)\alpha(p-1))} - (y_t - \exp(\mathbf{x}_t\beta + \mathbf{H}_t(p-1)\alpha(p-1)))\right)$$

와 같이 정의될 수 있다. 이 때 $H_t(p) = (H_{t-1}, \cdots, H_{t-p})$이며, $\alpha(p) = (\alpha_1, \cdots, \alpha_p)^T$ 이다. 이와 같이 편차의 차로 표현된 통계량은 $\chi^2(p_A - p_B)$ 분포를 따르며(여기서는 $\chi^2(1)$이 된다.) $H_0 : \alpha_p = 0$를 검정하기 위한 것이다. 모형 2의 경우도 마찬가지로 적용해 볼 수 있겠다.

특별히 자료가 과대 산포나 과소 산포되어 있을 경우, 척도화 편차(scaled deviance)를 사용한다. 이는 편차를 산포의 척도인 산포모수(dispersion parameter) ϕ로 나누어 준 값이다. 위의 예의 경우에 척도화 편차의 차는 다음과 같이 표현된다.

$$\frac{D(y;\hat{\tau}_{At}|\mathbf{x}_t,(H_{t-1},\cdots,H_{t-p+1})) - D(y;\hat{\tau}_{Bt}|\mathbf{x}_t,(H_{t-1},...,H_{t-p}))}{\phi} \tag{3.3}$$

이는 근사적으로 $\chi^2(1)$을 따른다.

만약 산포모수를 모를 경우에는 다음과 같은 추정치를 사용한다.

$$\hat{\phi} = \frac{1}{n-p}\sum D(y_t;\hat{\tau}_t|\mathbf{x}_t,\mathbf{H}_t).$$

4. Polio 자료 분석

자기상관성이 있는 비정규 시계열의 분석을 위해 지금까지 제안된 확장된 일반화 선형 모델을 응용하고 비교해 보기 위하여 1970년부터 1980년까지의 소아마비 자료를 분석하고자 한다. 이 자료는 미국의 질병 예방 센터(U.S. Centers for Disease Control)에 의해서 보고된 1970년부터 1980년까지 매달 소아마비 숫자이다. 우리의 주 관심사는 소아마비의 숫자가 10년 동안 통계적으로 유의하게 감소했다고 볼 수 있는가에 대한 문제이다. 주된 설명변수로 Trend=0,1,2,....167 를 취하였고, 이는 한 달 간격의 시간을 의미한다. 시계열 자료의 계절성을 감안해서 매년 그리고 반년 주기인 sine, cosine함수를 설명변수로 하였다.

자기상관성을 가지는 비정규 시계열 자료가 포아송 분포를 따른다고 가정했을 때, 가능한 모델을 살펴 보면 다음과 같다.

표 2. 적합 모형의 종류

모형1-1	$\mathbf{H}_t = (y_{t-1} - \tau_{t-1}, \cdots, y_{t-p} - \tau_{t-p})$인 조건부 모형
모형1-2	$\mathbf{H}_t = (y_{t-1} - \exp(\mathbf{X}_{t-1}\beta), \cdots, y_{t-p} - \exp(\mathbf{X}_{t-p}\beta))$인 조건부 모형
모형1-3	$\mathbf{H}_t = (y_{t-1}, \cdots, y_{t-p})$인 조건부 모형
모형1-4	$\mathbf{H}_t = (\log(y_{t-1}) - \mathbf{X}_{t-1}\beta, \cdots, \log(y_{t-p}) - \mathbf{X}_{t-p}\beta)$ 인 조건부 모형
모형2	AR(p)오차를 이용한 조건부 모형
모형3	주변 모형

소아마비 시계열 자료를 위의 6가지 모형에 실제로 적합 시켜 보았다. 관측치에 의한 모형에 대해서는 모형의 시차(lag)를 1부터 5까지 변화시키면서 모두 적합하였다. 또한 Lee(2001)을 참조하여 전체 자료의 개수가 n 개일 때 $n^{1/3}$ 정도까지의 차수를 비교하였다. 다음 표 1,2 는 위의 모형 중 조건부 모형에 대해 3장에서 설명한 AIC_{Diff}와 BIC_{Diff}를 계산한 결과이다.

표 3. 조건부 모형의 AIC_{Diff}

AIC_{Diff}	모형 1-1	모형 1-2	모형 1-3	모형 1-4	모형 2
시차(lag) 0	266.213	266.213	266.213	266.213	266.213
$\alpha_1 H_{t-1}$	257.661	257.392	257.301	256.620	253.241
$\alpha_1 H_{t-1} + \alpha_2 H_{t-2}$	257.702	258.412	258.370	250.236	251.902
$\alpha_1 H_{t-1} + \cdots + \alpha_3 H_{t-3}$	259.484	259.994	260.011	240.891	251.022*
$\alpha_1 H_{t-1} + \cdots + \alpha_4 H_{t-4}$	261.509	260.348	260.276	241.528	252.167
$\alpha_1 H_{t-1} + \cdots + \alpha_5 H_{t-5}$	257.619*	256.631*	257.073*	239.341*	254.183

표 4. 조건부 모형의 BIC_{Diff}

BIC_{Diff}	모형 1-1	모형 1-2	모형 1-3	모형 1-4	모형 2
시차(lag) 0	284.776	284.776	284.776	284.776	284.776
$\alpha_1 H_{t-1}$	279.317*	279.048*	278.958*	278.277	274.903*
$\alpha_1 H_{t-1} + \alpha_2 H_{t-2}$	282.452	283.162	283.120	274.986	276.666
$\alpha_1 H_{t-1} + \cdots + \alpha_3 H_{t-3}$	287.328	287.838	287.855	268.735*	278.865
$\alpha_1 H_{t-1} + \cdots + \alpha_4 H_{t-4}$	292.447	291.286	291.214	272.466	283.104
$\alpha_1 H_{t-1} + \cdots + \alpha_5 H_{t-5}$	291.651	290.662	291.104	273.372	288.214

차수가 0인 경우는 보통의 일반화 선형 모형이다. 이 것은 독립성을 가정하고 있어 관측값의 자기상관성을 설명해주지 못한다. 전체 자료의 개수 n에 대해 $n^{1/3} \simeq 5$이므로, 각각의 모형마다 차수가 1부터 5까지 적합하였고 이들 모형은 과거 관측치를 설명변수로 하여 자료의 상관관계를 설명해주고 있다. 모형 2을 제외하고는 모두 AIC_{Diff}값이 시차 (lag)가 5일 때 가장 작다. 모형 2의 경우에는 시차(lag) 3인 경우가 최소값을 갖는다. 따라서 AIC_{Diff}를 기준으로 하면 모형 2을 제외한 각각의 모형에서는 시차(lag)가 5일 때가 가장 좋으며, 모형 2의 경우는 시차 (lag) 3일 때가 가장 적합하다. 그러나 BIC_{Diff}의 경우는 결과가 좀 다르게 나타난다. 모형 1-4를 제외한 모든 모형은 시차 (lag)가 1일 때 가장 작으며, 모형 1-4에서는 시차 (lag) 3인 경우가 가장 작았다. 서로 다른 모형의 비교를 보면, 모형 1-4가 전반적으로 다른 모형들에 비해 AIC_{Diff}와 BIC_{Diff}값이 작은 것을 알 수 있다. 이는 모형 1-4가 다른 모형들에 비해 자료에 더 잘 적합하다는 것을 의미한다. 최종적으로 AIC_{Diff}는 시차 (lag)가 5인 모형 1-4를, BIC_{Diff}는 시차 (lag)가 3인 모형 1-4를 각각 가장 적합한 모형임을 보여준다.

다음 표 5,6,7는 편차(deviance)의 차를 이용한 모형 1-2,1-3,1-4에 대한 분석 결과이다.

표 5. 모형 1-2의 편차 분석표

모형 1-2	편차	df	편차의 차	df	P-값	척도화 편차의 차	p-값
시차 (lag) 0	281.715	157	-	-	-	-	-
$\alpha_1 H_{t-1}$	271.174	156	10.541	1	0.001*	6.229	0.013*
$\alpha_1 H_{t-1} + \alpha_2 H_{t-2}$	270.187	155	0.987	1	0.320	0.583	0.445
$\alpha_1 H_{t-1} + \cdots + \alpha_3 H_{t-3}$	269.830	154	0.357	1	0.550	0.211	0.646
$\alpha_1 H_{t-1} + \cdots + \alpha_4 H_{t-4}$	268.236	153	1.594	1	0.207	0.942	0.332
$\alpha_1 H_{t-1} + \cdots + \alpha_5 H_{t-5}$	262.513	152	5.723	1	0.017*	3.382	0.066

모형 1-2의 산포모수의 추정치는 $\hat{\phi} = 1.692$이다.

표 6. 모형 1-3의 편차 분석표

모형 1-3	편차	df	편차의 차	df	P-값	척도화 편차의 차	p-값
시차 (lag) 0	281.715	157	-	-	-	-	-
$\alpha_1 H_{t-1}$	271.055	156	10.660	1	0.001*	6.266	0.012*
$\alpha_1 H_{t-1} + \alpha_2 H_{t-2}$	270.105	155	0.950	1	0.330	0.558	0.455
$\alpha_1 H_{t-1} + \cdots + \alpha_3 H_{t-3}$	269.816	154	0.289	1	0.591	0.170	0.680
$\alpha_1 H_{t-1} + \cdots + \alpha_4 H_{t-4}$	268.156	153	1.660	1	0.197	0.976	0.323
$\alpha_1 H_{t-1} + \cdots + \alpha_5 H_{t-5}$	262.956	152	5.200	1	0.023*	3.057	0.080

모형 1-3의 산포모수의 추정치는 $\hat{\phi} = 1.701$이다.

표 7. 모형 1-4의 편차 분석표

모형 1-4	편차	df	편차의 차	df	P-값	척도화 편차의 차	p-값
시차 (lag) 0	281.715	157	-	-	-	-	-
$\alpha_1 H_{t-1}$	270.423	156	11.292	1	0.001*	7.353	0.007*
$\alpha_1 H_{t-1} + \alpha_2 H_{t-2}$	262.040	155	8.383	1	0.004*	5.459	0.019*
$\alpha_1 H_{t-1} + \cdots + \alpha_3 H_{t-3}$	250.725	154	11.315	1	0.001*	7.368	0.007*
$\alpha_1 H_{t-1} + \cdots + \alpha_4 H_{t-4}$	249.400	153	1.325	1	0.250	0.863	0.353
$\alpha_1 H_{t-1} + \cdots + \alpha_5 H_{t-5}$	245.223	152	4.177	1	0.041*	2.720	0.100

모형 1-4의 산포모수의 추정치는 $\hat{\phi} = 1.536$이다.

표 5, 6, 7에서의 척도화 편차의 차는 3절에서 설명한 바와 같이 편차를 산포모수 (dispersion parameter)로 나눈 값이다. 만약 자료가 과대산포 또는 과소산포되어 있는 경우에는 척도화 편차의 차를 통해 비교하는 것이 바람직하다. 척도화 편차의 차를 통해 각 모형 비교한 결과, 모형 1-2, 1-3에서는 시차 (lag) 1인 경우가 가장 적합하고, 모형 1-4에서는 시차 (lag) 3까지가 유의한 것으로 나타났다. 이 결과는 BIC_{Diff}에서의 결과와 동일한 결과이며, 시차(lag) 3의 모형 1-4 가 BIC_{Diff}값이 268.735로 모든 모형 중 가장 작으므로 관측치에 의한 모형 중 가장 좋은 모형으로 선택 할 수 있겠다.

다음 표 8 은 보통의 일반화 선형 모형과 조건부 모형 중 가장 좋은 모형으로 선택한 시차 (lag) 3인 모형 1-4, 그리고 주변 모형의 AIC_{Diff}와 BIC_{Diff}값을 비교한 것이다. 그리고 표 9는 각각의 모형에 대한 계수의 추정치와 표준 오차를 나타낸 표이다.

표 8. AIC_{diff}와 BIC_{Diff} 으로의 모형 비교

	일반화 선형모형	조건부 모형(모형 1-4, lag 3)	주변 모형
AIC_{Diff}	266.213	240.891*	294.899
BIC_{Diff}	284.776	268.735*	313.461

표 9. 소아마비 시계열자료 의 분석 결과

	GLM		조건부 모형	(모형 5)	주변 모형	
설명변수	$\hat{\beta}$	표준오차	$\hat{\beta}, \hat{\alpha}$	표준오차	$\hat{\beta}, \hat{\sigma}, \hat{\rho}$	표준오차
$Trend \times 10^{-3}$	-4.80	1.95	-4.51	1.78	-4.35	2.68
$\cos(2\pi t/12)$	-0.15	0.14	-0.14	0.12	-0.11	0.16
$\sin(2\pi t/12)$	-0.53	0.15	-0.49	0.14	-0.48	0.17
$\cos(2\pi t/6)$	0.17	0.14	0.16	0.12	0.20	0.14
$\sin(2\pi t/6)$	-0.43	0.14	-0.42	0.13	-0.41	0.14
H_{t-1}			0.12	0.04	$\hat{\sigma}^2$	0.77
H_{t-2}			0.08	0.04	$\hat{\rho}_y(1)$	0.25
H_{t-3}			-0.09	0.03	$\hat{\rho}_\epsilon(1)$	0.77
$\hat{\phi}$	1.95		1.56		1.0	

표 8에서와 같이 서로 다른 모형들간의 객관적 비교가 가능하며, 그 결과 세 모형 중 조건부 모형이 가장 적합한 모형임을 알 수 있다. 주된 관심사인 Trend에 대한 계수의 추정치는 조건부 모형의 경우 -4.51이며, 이는 일반화 선형모형의 -4.80과 모수에 의한 모형의 -4.35의 중간값을 나타내었다. 앞서 주변 모형을 제시한 Zeger (1988)는 주변 모형를 사용한 결과를 이용하여, 계수의 추정치 -4.35 그리고 표준오차 2.68로서 소아마비 숫자가 10년 동안 감소했다고 할 수 없다는 결론을 지었다. 그러나 우리가 가장 적합한 모형으로 선택한 조건부 모형의 경우, Trend 계수의 추정치에 대한 표준오차가 1.78로 나머지 두 모형에 비해 작은 값을 가지며, 따라서 장기간 소아마비 숫자가 줄고 있다는 증거가 있다고 볼 수 있겠다. 이는 예전 Zeger (1988)의 결과와는 정반대의 결과를 지지해 주고 있으며, 이는 어떤 모형을 선택하는냐에 따라 분석의 결과는 달라질 수 있다는 것을 보여 준다. 따라서 모형 선택은 매우 중요한 문제라 할 수 있겠다.

5. 맺음말

본 논문은 지금까지 비정규 시계열 자료에 대해 개별적으로 적합된 회귀 모형들을 통합하여 관련된 모든 가능한 모형들을 소개하고 그 차이점을 알아보았다. 또한 모형선택의 기준으로 기존의 AIC, BIC, 우도비 검정통계량들을 확장 적용하였고, 이러한 내용을 실제의 자료분석에 응용하여 분석결과를 제시해 보았다. 우리가 분석한 소아마비 자료의 경우, AIC는 시차 (lag)가 5인 모형 1-4를, BIC는 시차 (lag)가 3인 모형 1-4를 각각 가장 적합한 모형으로 선택하였다. 그러나 척도화 편차의 차를 이용하여 비교한 결과 BIC와 동일한 결과가 나왔으므로, 우리는 시차 (lag)가 3인 모형 1-4를 가장 적합한 모형으로 선택하였다. 자료의 특징에 따라 AIC, BIC 또는 편차의 차 등 어떤 기준을 이용하는가는 다소 달라 질

수 있겠다. 또한 서로 다른 모형들간의 비교, 즉 일반화 선형모형, 조건부 모형, 그리고 주변 모형간의 비교를 AIC와 BIC를 통해 해 보았다. 이는 각 모형들간의 객관적 비교를 가능하게 하였으며, 그 결과 조건부 모형이 Polio자료에 가장 적합함을 알 수 있었다. 본문에서 설명한 것과 같이 모형 1-4를 이용한 결과는 예전 Zeger (1988) 이 주변 모형을 통해 분석한 결과와는 정반대의 결과를 지지하고 있었다. 이를 통해 우리는 가장 적합한 모형의 객관적 선택이 상당히 중요한 문제라는 것을 인식할 수 있다. 또한 모형들간의 비교에서 그치지 않고, 더 나아가 회귀진단에 대한 연구가 이루어진다면 보다 정확하고 객관적인 분석이 가능할 것이라 생각된다.

참고문헌

[1] Akaike, H.(1969). Fitting autoregressive models for prediction. *Ann. Inst. Statist. Math.* **21** 243-247.

[2] Akaike, H.(1977). In entropy maximisation principle. In*Application of Statistics* (Krishnaiah,P.R. ed.) 27-41. North Halland, Amsterdam.

[3] Akaike,H.(1979). A Bayesian Extension of the Minimum AIC Procedure of Autoregressive Model Fitting . *Biometrika* **66**, 237-242.

[4] Brumback,B., Ryan,L. Schwartz,J.,Neas,L., Stark,P., and Burge,H.(2000). Transitional Regression Models, with Application to Environmental Time Series.*Journal of the American Statistical Assoc.* **95**, 16-27.

[5] Cox,D.R.(1981). Statistical Analysis of Time Series: Some Recent Developments. *Scandinavian Journal of Statistics* **8**, 93-115.

[6] Fahrmeir,L. and Tutz,G.(2001). *Multivariate Statistical Modelling Based on Generalized Linear Model* . 2nd edition, New York: Springer-Verlag.

[7] Hastie,T.J. and Tibshirani, R.J.(1990), *Generalized Additive Models.* London: Chapman and Hall.

[8] Korn, E.L. and Whittemore, A.S. (1979). Methods for analyzing panel studies of acute health effects of air pollution. Biometrics 35, 795-802.

[9] Kalbfleisch, J.D. and Lawless, J.F.(1985). The analysis of panel data under a Markov assumption. Journal of the American Statistical Association 80, 863-871.

[10] Lee,S. and Karagrigoriou,A.(2001). An Asymptotically Optimal Selection of the Order of a Linear Process. *The Indian journal of Statistics* **63**, Series A, Pt. 1, 93-106.

[11] Li,W.K.(1994). Time Series Models Based on Generalized Linear Models: Some Further

Results. *Biometrics* **50**, 506-511.

[12] Liang,K.Y. and Zeger,S.L.(1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13-22.

[13] McCullagh,P. and Nelder,J.A.(1989) *Generalized Linear Models.* 2nd edition, London: Chpman and Hall.

[14] Pawitan,Y.(2001). *In All Likelihood - Statistical Modelling and Inference Using Likelihood.* Oxford University Press.

[15] Prentice,R.L.(1988). Correlated Binary Regression with Covariates Specific to Each Binary Observation. *Biometrics* **44**, 1033-1048.

[16] Reinsel,G.(1997). *Elements of Multivariate Time series Analysis.* 2nd edition, New York: Springer-Verlag.

[17] Samet,J.M., Zeger,S.L. and Berhane,K.(1995). "Then Association of mortality and particulate Air Pollution" in *Particulate Air Pollution and Daily Mortality; Replication and Validation of Selected Studies*, The phase I Report of the Particle Epidemiology Evaluation Project, Health Effects Institute.

[18] Schwarz,G.(1978).Estimating the Dimension of a Model. *Annals of Statistics* **6**, 461-464.

[19] Stone,M.(1977). An Asymptotic Equicalence of Choice of Model by Cross-Validation and Akaike's Criterion. *Journal of the Royal Statistical Society. Series B(Methodological)* **39**, 44-47.

[20] Zeger,S.L.(1988). A Regression model for time series of counts.*Biometrika* **75**, 621-629.

[21] Zeger,S.L. and Qaqish,B.(1988). Markov Regression models for Time Series: A Quasi-Likelihood Approach. *Biometrics* **44**, 1019-1031.

[2003년 4월 접수, 2003년 6월 채택]

Generalized Linear Model with Time Series Data

Yoonha Choi [1] Sungim Lee [2] S. Lee [3]

ABSTRACT

In this paper we reviewed a variety of non-Gaussian time series models, and studied the model selection criteria such as AIC and BIC to select proper models. We also considered the likelihood ratio test and applied it to analysis of Polio data set.

Keywords: Akaike information criterion; Bayesian informaition criterio; deviance; non-Gaussian time series; regression model; model selection.

1) M.A., Department of Statistics, Seoul National University.
E-mail : yhchoi@stats.snu.ac.kr
2) Assistant research professor,Medical Science Research Center, Korea University.
E-mail : silee70@ns.kumc.or.kr
3) Professor, Department of Statistics, Seoul National University.
E-mail : sylee@stats.snu.ac.kr