The Korean Journal of Applied Statistics (2015) DOI: http://dx.doi.org/10.5351/KJAS.2015.28.3.583
28(3), 583–592

Zero-Inflated INGARCH Using Conditional Poisson and Negative Binomial: Data Application

J. E. Yoon[a] · S. Y. Hwang[a,1]

[a]Department of Statistics, Sookmyung Women's University

(Received May 26, 2015; Revised June 8, 2015; Accepted June 9, 2015)

Abstract

Zero-inflation has recently attracted much attention in integer-valued time series. This article deals with conditional variance (volatility) modeling for the zero-inflated count time series. We incorporate zero-inflation property into integer-valued GARCH (INGARCH) via conditional Poisson and negative binomial marginals. The Cholera frequency time series is analyzed as a data application. Estimation is carried out using EM-algorithm as suggested by Zhu (2012).

Keywords: integer-valued time series, conditional Poisson, zero-inflated INGARCH

1. 서론

정수 값을 갖는(integer valued) 계수 시계열(count time series)은 다양한 분야에서 접할 수 있으며 계수 시계열의 특성을 반영한 효과적인 분석을 위한 연구가 활발히 이루어지고 있다. 계수 시계열의 일차 적률인 조건부 평균(conditional mean)을 분석하기 위한 표준적인 모형은 INAR(Integer-valued AR) 모형 (Hwang과 Basawa, 2011)이며 Grunwald 등 (2000)은 표준적인 binomial thinning을 일반화 시킨 다양한 INAR 모형의 변형에 대해 연구하였다. 전통적인 조건부 포아송, 조건부 멱급수 분포 및 조건부 자기로지스틱 모형 등도 계수 시계열의 조건부 평균을 분석하는 도구이다. 계수 시계열의 변동성(volatility)을 분석하기 위한 모형으로는 INGARCH 모형이라 불리는 조건부 이분산성(conditionally heteroscedastic)을 고려한 INteger-valued GARCH (Ferland 등, 2006) 모형이 있다. 이 모형은 대표적인 이분산 시계열 모형인 GARCH 모형과 유사한 수학적 성질을 갖는 모형으로 조건부 평균 모형인 INAR을 조건부 이분산 모형으로 확장시킨 모형이라 할 수 있다 (Yoon과 Hwang, 2015). 조건부 분포로 흔히 이용하는 포아송(Poisson) 주변분포로는 계수 시계열의 변동성 분석에서 흔히 발생하는 과산포(over-dispersion) 문제를 다루기 어렵다는 단점이 있다고 알려져 있다. 이를 해결하기 위하여 Zhu (2011, 2012)는 조건부 분포로 음이항분포를 이용한 Negative Binomial INGARCH(즉, NBINGARCH) 모형을 제안하였다. 계수형 국내 자료의 INGARCH 모형들의 사례분석은 Yoon과 Hwang (2015)을 참고하기 바란다.

본 논문에서는 영-과잉 현상(zero-inflation)이 존재하는 영-과잉 계수 시계열(zero-inflated count time series)의 변동성을 연구하고자 한다. 계수 시계열 자료가 특정한 질병의 발병, 특정한 범죄의 발생 등

[1]Corresponding author: Department of Statistics, Sookmyung Women's University, Yongsan-Gu, Seoul 140-742, Korea. E-mail: shwang@sookmyung.ac.kr

일반적으로 흔히 일어나지 않는 사건들에 대해 조사하는 경우에는 많은 수의 0이 관측되는 경우가 발생한다. 계수 시계열 자료에 대한 실증자료 분석에서 과산포와 더불어 발생하는 문제가 이와 같은 영-과잉 현상이다. 영-과잉 현상이란 포아송 모형을 가정하는 경우 모형에서 기대되는 0값보다 더 많은 수의 0이 관측되는 현상을 말한다 (Zhu, 2012). 이러한 경우 기존의 INGARCH 모형 또는 NBINGARCH 모형을 그대로 적합 시킨다면 모형의 적합도가 떨어질 것이며, 과산포 문제를 심화시킬 수 있을 것이다. 이러한 영-과잉 문제를 해결하는 방안으로 Zhu (2012)는 자료에 포함된 0에 대해 적절한 가중치를 주어 다루는 영-과잉 변동성 모형으로서 조건부 분포로 기존의 포아송분포 대신 영-과잉 포아송분포(Zero-inflated Poisson; ZIP)로 대체한 ZIP-INGARCH 모형, 음이항분포 대신 영-과잉 음이항분포(Zero-inflated Negative Binomial; ZINB)로 대체한 ZINB-INGARCH 모형을 제안하였다. 본 논문에서는 영-과잉 계수 시계열의 변동성을 효과적으로 분석하기 위해 제안된 다양한 영-과잉 계수형 GARCH 모형들을 소개하고 이러한 모형에서 모수를 추정하는 방법으로 EM 방법에 대해 단계별로 자세히 제시하고 있으며 제안된 방법론을 국내 콜레라 발생건수 자료에 적용하여 보았다.

2. 영-과잉 계수형 GARCH 모형: ZIP-INGARCH 모형과 ZINB-INGARCH 모형

영-과잉 현상에서 0은 두 가지로 생각할 수 있다. 예를 들어, 특정 바이러스 감염자 수 자료를 생각해보자. 조사 대상자가 그 바이러스에 대해 저항력이 있어 바이러스에 노출은 되었으나 감염이 되지 않았을 수 있고(structural zero), 또는 그 바이러스에 아예 노출되지 않아서 감염이 되지 않았을 수도 있다(sampling zeros). 이러한 경우 각각 예상할 수 있는 0과 우연히 발생하는 0으로 볼 수 있다. 영-과잉 자료에 있는 0을 다루기 위해서는 이러한 두 경우를 모두 고려하여 적절히 처리하는 것이 필요하다.

계수 시계열 자료에서 영-과잉 자료인지를 파악하는 방법으로 널리 이용되는 Puig와 Valero (2006)가 제안한 zero-inflation index 즉, ZI index는 다음과 같이 계산한다.

$$\text{ZI index} = 1 + \frac{\log(p_0)}{\mu},$$

여기서 p_0는 0값의 비율(proportion)이고 μ는 평균이다. ZI index가 0보다 크면 영-과잉 자료로 판단한다 (Puig와 Valero, 2006). 영-과잉 자료로 판단되는 경우 영-과잉 변동성 모형을 적합 하는 것이 필요하다. 영-과잉 변동성 모형들의 소개는 Zhu (2011, 2012)와 Yoon과 Hwang (2015)을 참고하였다.

2.1. ZIP-INGARCH(p, q) 모형

계수형 시계열 자료를 다루기 위한 모형으로 Ferland 등 (2006)이 제안한 integer-valued GARCH, INGARCH(p, q) 모형은 다음과 같다.

$$\begin{aligned} X_t|F_{t-1} &: \text{Poisson}(\lambda_t), \\ \lambda_t &= \alpha_0 + \sum_{i=1}^{p} \alpha_i X_{t-i} + \sum_{j=1}^{q} \beta_j \lambda_{t-j}, \end{aligned}$$

여기서 $\alpha_0 > 0$, $\alpha_i \geq 0$, $\beta_j \geq 0$, $i = 1, \ldots, p$, $j = 1, \ldots, q$, $p \geq 1$, $q \geq 0$이다. 여기서 $\{X_t\}$는 각 자료를 나타내고 F_{t-1}은 $t-1$시점까지의 정보 집합이다. 이 모형은 조건부 분포로 포아송분포를 가정하고 있다. 평균과 분산이 같다는 특징을 가진 포아송 분포의 특성상 이 모형에서는 계수 시계열 자료의 분석에서 흔히 발생하는 과산포(over-dispersion) 문제를 설명하지 못한다는 단점이 있었다. 이러한 과산포 문제와 영-과잉 현상을 동시에 효과적으로 다루기 위해 Zhu (2012)는 다음과 같은 ZIP-INGARCH(p, q)

모형을 제안하였다.

$$X_t|F_{t-1} : \text{ZIP}(\lambda_t, w),$$
$$\lambda_t = \alpha_0 \ + \sum_{i=1}^{p} \alpha_i X_{t-i} + \sum_{j=1}^{q} \beta_j \lambda_{t-j},$$

여기서 $0 < w < 1$, $\alpha_0 > 0$, $\alpha_i \geq 0$, $\beta_j \geq 0$, $i = 1, \ldots, p$, $j = 1, \ldots, q$, $p \geq 1$, $q \geq 0$이다.

ZIP-INGARCH 모형이 기존의 INGARCH 모형과 다른 점은 조건부 분포로 포아송 분포 대신 영-과잉 포아송분포(Zero-inflated Poisson; ZIP)를 가정했다는 점이다. ZIP(λ, w)는 영-과잉 포아송 분포의 확률질량함수로 다음과 같다.

$$P(X = k) = w\delta_{k,0} + (1 - w)\frac{\lambda^k e^{-\lambda}}{k!}, \quad k = 0, 1, 2 \ldots,$$

$0 < w < 1$이고 $\delta_{k,0}$는 $k = 0$인 경우에는 1이 되고, $k \neq 0$인 경우에는 0이 되어 0의 확률을 구하는 데 영향을 준다. $w = 0$이라면 이 모형은 영-과잉 현상을 반영하지 않은 INGARCH(p, q) 모형이 된다. ZIP-INGARCH 모형에서 조건부 평균과 조건부 분산은 다음과 같이 유도된다.

$$E(X_t|F_{t-1}) = (1 - w)\lambda_t,$$
$$\text{Var}(X_t|F_{t-1}) = (1 - w)\lambda_t(1 + w\lambda_t).$$

위의 식을 보면 조건부 분산이 조건부 평균보다 더 큰 값을 가짐을 알 수 있다. 이를 통해 영-과잉 현상 뿐 아니라 INGARCH(p,q) 모형에서 설명하기 어려운 과산포 문제 또한 다룰 수 있게 되었다. ZIP-INGARCH(p, q) 모형에서 $q = 0$이면 ZIP-INARCH(p) 모형이 되며, 차수가 1인 ZIP-INARCH(1) 모형은 다음과 같다 (Zhu, 2012).

$$X_t|F_{t-1} : \text{ZIP}(\lambda_t, w),$$
$$\lambda_t = \alpha_0 \ + \alpha_1 X_{t-1},$$

여기서 $\alpha_0 > 0$이며 정상성(stationarity)을 만족하기 위해 $(1 - w)\alpha_1^2$은 1보다 작다.

2.2. ZINB-INGARCH(p, q) 모형

INGARCH 모형에서 과산포 문제를 다루기 어렵다는 단점을 보완하기 위하여 Zhu (2011)가 제안한 음이항 INGARCH, 즉, NBINGARCH(p, q) 모형은 다음과 같다.

$$X_t|F_{t-1} : NB(r, p_t),$$
$$\lambda_t = \frac{1 - p_t}{p_t} = \ \alpha_0 \ + \sum_{i=1}^{p} \alpha_i X_{t-i} + \sum_{j=1}^{q} \beta_j \lambda_{t-j},$$

여기서 NB는 음이항분포(negative binomial distribution)이며 r은 양의 정수이고 $\alpha_0 > 0$, $\alpha_i \geq 0$, $\beta_j \geq 0$, $i = 1, \ldots, p$, $j = 1, \ldots, q$, $p \geq 1$, $q \geq 0$이다. 조건부 분포를 음이항분포로 가정함으로 포아송 분포로는 설명이 어려운 과산포 문제와 극단적인 관측값을 효과적으로 다룰 수 있는 모형이다.

이 모형에서 과산포 문제 뿐 아니라 영-과잉 현상을 효과적으로 다루기 위해 조건부 분포로 영-과잉 음이항분포(Zero-inflated Negative Binomial; ZINB)를 이용한 ZINB-INGARCH(p, q) 모형은 다음과 같

다.

$$X_t F_{t-1} : \text{ZINB}(\lambda_t, a, w),$$
$$\lambda_t = \frac{1-p_t}{p_t} = \alpha_0 + \sum_{i=1}^{p} \alpha_i X_{t-i} + \sum_{j=1}^{q} \beta_j \lambda_{t-j},$$

$0 < w < 1$, $a > 0$, $\alpha_0 > 0$, $\alpha_i \geq 0$, $\beta_j \geq 0$, $i = 1, \ldots, p$, $j = 1, \ldots, q$, $p \geq 1$, $q \geq 0$이고, 여기서 $\text{ZINB}(\lambda, a, w)$는 영-과잉 음이항분포의 확률질량함수로 다음과 같이 정의된다.

$$P(X=k) = w\delta_{k,0} + (1-w)\frac{\Gamma\left(k + \frac{\lambda^{1-c}}{a}\right)}{k!\Gamma\left(\frac{\lambda^{1-c}}{a}\right)}\left(\frac{1}{1+a\lambda^c}\right)^{\frac{\lambda^{1-c}}{a}}\left(\frac{a\lambda^c}{1+a\lambda^c}\right)^k, \quad k = 0, 1, 2, \ldots,$$

여기서 $\lambda > 0$, $0 < w < 1$, $a > 0$이며 $\delta_{k,0}$은 $k = 0$인 경우에는 1이 되고, $k \neq 0$인 경우에는 0이 되어 0의 확률을 구하는데 영향을 준다. 또한 a는 퍼진 정도를 나타내는 모수로 0보다 크거나 같은 값을 가지며, a값이 0으로 간다면 이 분포는 영-과잉 포아송분포가 된다. c값을 강조하지 않은 경우 일반적으로 $\text{ZINB}(\lambda, a, w)$로 표기하나 $c = 0$인 경우에는 $\text{ZINB1}(\lambda, a, w)$로, $c = 1$인 경우에는 $\text{ZINB2}(\lambda, a, w)$로 구분하여 나타내기도 한다.

조건부 평균과 조건부 분산은 다음과 같이 유도된다.

$$E(X_t|F_{t-1}) = (1-w)\lambda_t,$$
$$\text{Var}(X_t|F_{t-1}) = (1-w)\lambda_t(1 + w\lambda_t + a\lambda_t^c).$$

위의 식을 보면 조건부 분산이 조건부 평균보다 더 큰 값을 가지므로, 과산포 문제를 또한 다룰 수 있음을 알 수 있다. ZINB-INGARCH(p, q) 모형에서 $q = 0$이면 ZINB-INARCH(p) 모형이 되며, 차수가 1인 ZINB-INARCH(1) 모형은 다음과 같다.

$$X_t|F_{t-1} : \text{ZINB}(\lambda_t, a, w),$$
$$\lambda_t = \frac{1-p_t}{p_t} = \alpha_0 + \alpha_1 X_{t-1},$$

여기서 $\alpha_0 > 0$이고 정상성(stationarity)을 만족하기 위해 $(1-w)\alpha_1^2$은 1보다 작다.

3. 모수추정

영-과잉 계수형 GARCH모형에서 모수를 추정하기 위한 방법으로 EM알고리즘을 이용하였다. 이 방법은 Zhu (2012)가 제안한 방법으로 해당 논문에서는 방법론만을 제시하였으나 본 연구를 통해 프로그램을 작성한 후 다음 절의 사례분석에 이용하였다.

3.1. ZIP-INGARCH(p, q) 모형

ZIP-INGARCH(p, q) 모형의 가능도함수(likelihood function)는 다음과 같고

$$\prod_{X_t=0}\left[w + (1-w)e^{-\lambda_t}\right] \times \prod_{X_t>0}\left[(1-w)\frac{\lambda_t^{X_t}e^{-\lambda_t}}{X_t!}\right].$$

로그가능도함수(log-likelihood function)는 다음과 같다.

$$\sum_{X_t=0} \log\left[w + (1-w)e^{-\lambda_t}\right] + \sum_{X_t>0} \left[\log(1-w) + X_t \log\lambda_t - \lambda_t - \log(X_t!)\right].$$

위의 로그가능도함수를 최대로 하는 값을 찾음으로써 모수를 추정할 수 있으나, 이렇게 직접적으로 최대로 하는 값을 찾는 것은 상당히 복잡한 과정이 된다. 이러한 방법 대신 EM알고리즘을 이용하여 모수를 추정하는 방법을 소개하고자 한다.

$X = (X_1, \ldots, X_n)$을 ZIP-INGARCH(p, q) 모형에서 나온 자료로 가정한다. 새로운 변수 Z_t를 고려하여 자료에 대해 다음과 같이 생각하며, 특히 0값이 어떤 경우에 해당되는지 알고 있다고 가정한다.

- 경우 1 (난수발생 결과 생성된 0): $Z_t = 1$로 하며 $P(Z_t = 1) = w$.
- 경우 2 (포아송분포에서 생성): $Z_t = 0$으로 하며 $P(Z_0 = 0) = 1 - w$.

$Z = (Z_1, \ldots, Z_n)$, $\theta = (\alpha_0, \alpha_1, \ldots, \alpha_p, \beta_1, \ldots, \beta_q)^T = (\theta_0, \theta_1, \ldots, \theta_{p+q})^T$, $\Theta = (w, \theta^T)^T$이다. Θ에 대한 정보가 있는 경우 Z의 조건부분포, 즉 $(Z|\Theta)$의 분포는 다음과 같다.

$$\prod_{t=p+1}^{n} w^{Z_t}(1-w)^{1-Z_t}.$$

또한 Z, Θ에 대한 정보가 있는 경우 X의 조건부분포, 즉 $(X|Z, \Theta)$의 분포는 다음과 같다.

$$\prod_{t=p+1}^{n} \left(Z_t \times 1 + (1-Z_t) \times \frac{\lambda_t^{X_t} e^{-\lambda_t}}{X_t!}\right) = \prod_{t=p+1}^{n} \left(\frac{\lambda_t^{X_t} e^{-\lambda_t}}{X_t!}\right)^{1-Z_t}.$$

결과적으로, 전체자료에 대한 조건부가능도함수는 다음과 같이 유도되며

$$\prod_{t=p+1}^{n} w^{Z_t} \left((1-w)\frac{\lambda_t^{X_t} e^{-\lambda_t}}{X_t!}\right)^{1-Z_t},$$

전체자료에 대한 로그가능도함수는 다음과 같다.

$$l(\Theta) = \sum_{t=p+1}^{n} \{Z_t \log w + (1-Z_t)[\log(1-w) + X_t \log_{\lambda_t} - \lambda_t - \log(X_t!)]\}.$$

반복적인 EM알고리즘을 수행하여 위의 로그가능도함수를 최대로 만드는 추정값을 구한다. EM알고리즘은 다음과 같은 E단계와 M단계로 구성된다.

E단계: 추정하려고 하는 모수 Θ를 알고 있다고 가정한다. 결측값인 Z_t를 현재 알고 있는 모수를 이용하여 계산한 기댓값으로 대체하며, θ와 관측값의 정보를 이용하여 Z_t의 기댓값은 다음과 같이 계산된다.

$$\tau_t = \begin{cases} \dfrac{\omega}{\omega + (1-\omega)e^{-\lambda_t}}, & \text{if } X_t = 0, \\ 0, & \text{if } X_t = 1, 2, \ldots. \end{cases}$$

M단계: E단계에서 계산된 값을 이용하여 로그가능도함수를 새롭게 계산할 수 있다. 모수 Θ에 대한 추정은 로그가능도함수를 최대로 만드는 값으로 한다. M단계에서 이용되는 식은 다음과 같다.

$$\widehat{w} = \frac{1}{n-p}\sum_{t=p+1}^{n}\tau_t,$$
$$\sum_{t=p+1}^{n}(1-\tau_t)\left(\frac{X_t}{\lambda_t}-1\right)\frac{\partial\lambda_t}{\partial\theta_t}\Big|_{\widehat{\theta}} = 0, \quad i = 0, 1, \ldots, p+q.$$

위의 식은 닫힌 형태의 추정량(closed form)을 구할 수 없으므로 Newton-Raphson방법을 이용하여 추정값을 구한다. 초기값 $\theta^{(0)}$을 시작으로 다음 반복 시 업데이트되는 추정량은 다음과 같이 계산된다.

$$\theta^{(i+1)} = \theta^{(i)} - \left\{\frac{\partial^2 l}{\partial\theta\partial\theta^T}\Big|\theta^{(i)}\right\}^{-1}\frac{\partial l}{\partial\theta}\Big|_{\theta^{(i)}}.$$

$\theta^{(i)}$는 i번째 반복에서 추정되는 추정량이다. 이전 반복의 E단계에서 추정된 τ_t가 다음 단계의 Z_t의 값으로 사용된다. 모수 Θ에 대한 추정은 위의 두 단계를 수렴할 때까지 반복함으로 구한다. 수렴여부 판단 기준은 $|\,(\Theta_j^{(i+1)} - \Theta_j^{(i)})/\Theta_j^{(i)}\,| \le 10^{-5}$로 하였다.

3.2. ZINB-INGARCH(p,q) 모형

ZINB-INGARCH(p,q) 모형에서의 모수 추정방법도 앞의 ZIP-INGARCH(p,q) 모형의 모수 추정 방법과 유사하다. ZINB-INGARCH(p,q) 모형의 로그가능도함수(log-likelihood function)는 다음과 같다.

$$\sum_{x_t=0}\log\left[w + (1-w)(1+a\lambda_t^c)^{-\frac{\lambda_t^{1-c}}{a}}\right] + \sum_{x_t>0}\Bigg[\log(1-w) + X_t\log(a\lambda_t^c)$$
$$-\left(X_t + \frac{\lambda_t^{1-c}}{a}\right)\log\left(1+a\lambda_t^c\right) + \log\Gamma\left(X_t + \frac{\lambda_t^{1-c}}{a}\right) - \log\Gamma\left(\frac{\lambda_t^{1-c}}{a}\right) - \log(X_t!)\Bigg].$$

마찬가지로 위의 로그가능도함수를 최대로 하는 값을 찾음으로써 모수를 추정할 수 있으나 계산과정이 매우 복잡해지므로 EM알고리즘을 이용하여 모수를 추정하였다.

Z_t를 고려하여 만든 전체자료에 대한 조건부 로그가능도함수는 다음과 같다.

$$l(\Theta, a) = \sum_{t=p+1}^{n}(l_t^* + l_t^{**}),$$
$$l_t^* = Z_t\log\omega + (1-Z_t)\log(1-\omega),$$
$$l_t^{**} = (1-Z_t)\left[X_t\log\left(a\lambda_t^c\right) - \left(X_t + \frac{\lambda_t^c}{a}\right)\log\left(1+a\lambda_t^c\right) + \log\Gamma\left(X_t + \frac{\lambda_t^{1-c}}{a}\right)\right.$$
$$\left. - \log\Gamma\left(\frac{\lambda_t^{1-c}}{a}\right) - \log(X_t!)\right].$$

먼저 선택된 a값에 대하여 EM알고리즘을 수행하여 위의 조건부 로그가능도함수를 최대화하는 값을 추정한다. E단계에서 Z_t는 다음과 같이 추정된다.

$$\tau_t = \begin{cases} \dfrac{\omega}{\omega + (1-\omega)(1+a\lambda_t^c)^{-\frac{\lambda_t^{1-c}}{a}}}, & \text{if } X_t = 0, \\ 0, & \text{if } X_t = 1, 2, \ldots. \end{cases}$$

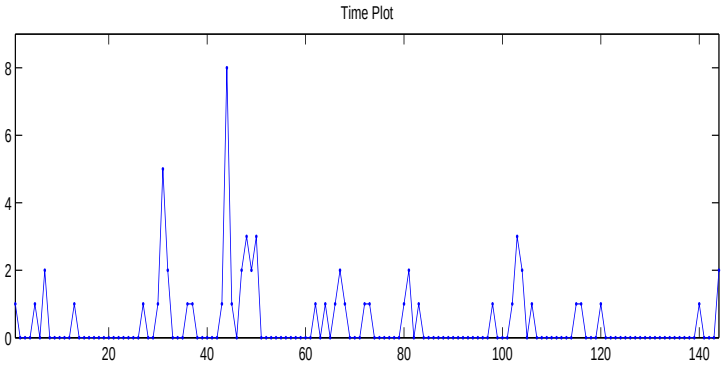

Figure 4.1. Count time series plot.

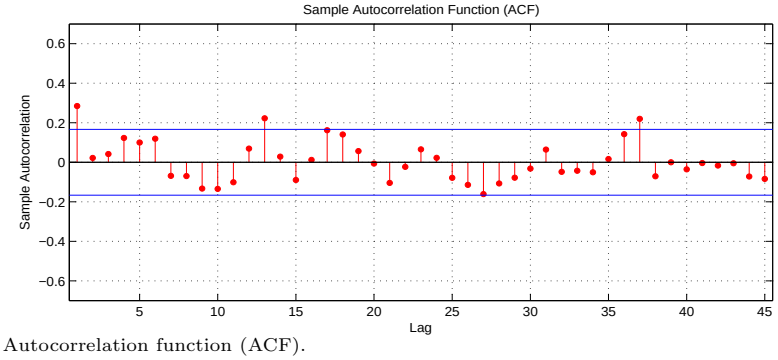

Figure 4.2. Autocorrelation function (ACF).

M단계에서는 앞의 모형에서와 동일한 과정을 거치며, 선택된 a값에 대하여 E단계와 M단계를 거쳐 $\widehat{\Theta}(a)$의 MLE를 구할 수 있다. 추정된 모수를 이용하여 a에 대한 profile likelihood인 $l(\widehat{\Theta}(a), a)$를 최대로 만들어주는 $\hat{a}$을 구할 수 있다. 다음 절의 사례분석에서는 $c = 0$인 ZINB1 모형과 $c = 1$인 ZINB2 모형에 대해 각각 모수를 추정하였다.

4. 사례분석

본 절에서는 2002년 1월부터 2013년 1월까지 보건복지부에 보고된 법정감염병발생보고 자료 중 콜레라 자료에 앞에서 소개한 모형들을 적합해 보았다. 이 자료는 매월 보고된 발생환자 수에 대한 기록으로 12년간 총 144개의 관측 값들로 구성된 계수 시계열 자료이다. Figure 4.1은 144개 자료에 대한 시도표이다. 시도표를 보면 관측된 0의 값이 많음을 알 수 있으며, 극단적으로 큰 값이 존재하기도 한다. 이 자료의 ZI index는 0.7004로 0보다 크므로 영-과잉 자료로 볼 수 있다. Figure 4.2와 Figure 4.3은 각각 자기상관함수(ACF)와 부분자기상관함수(PACF)에 대한 그래프이다. 자료의 평균과 분산은 각각 0.4306과 1.0161로 분산이 평균보다 크므로 과산포(over-dispersion)가 의심된다. 자료에 Poisson INARCH(1), NBINARCH(1), ZIP-INARCH(1), ZINB1-INARCH(1), ZINB2-INARCH(1) 모형을 적합시킨 결과는 Table 4.1과 같다. ZIP-INARCH(1), ZINB1-INARCH(1), ZINB2-INARCH(1) 모형의 모수 추정과정에서 w의 초기값은 0.5로 하였고, 각 모수의 초기값은 CLS 추정량을 이용하였다. a값의 초기값은 0.1로 주었다. 그 후 추정과정이 반복될 때마다 전단계의 a값을 초기값으로 사용하였으며,

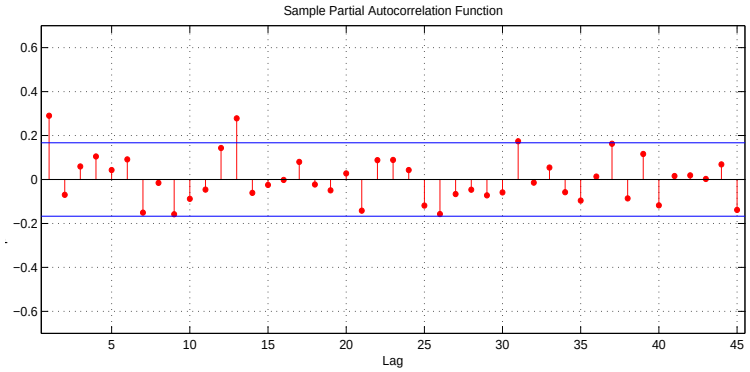

Figure 4.3. Partial autocorrelation function (PACF).

Table 4.1. Parameter estimates for INARCH(1), NBINARCH(1), ZIP-INARCH(1), ZINB-INARCH(1)

Model	$\hat{\omega}$	$\hat{\alpha_0}$	$\hat{\alpha_1}$	$\hat{a}$	AIC	BIC
Poisson		0.2413	0.4513		251.5690	257.4807
NB1		0.2851	0.3475	0.8059	235.3042	244.1717
NB2		0.2313	0.5401	0.9990	232.4747	241.3422
ZIP	0.4847	0.5169	0.5969		245.6278	254.4953
ZINB1	0.1183	0.3207	0.3886	0.7270	237.7142	249.5376
ZINB2	0.1842	0.2828	0.6579	0.9990	234.3636	246.1869

AIC값과 BIC값을 비교하여 이전 단계에서 변화가 없는 경우 과정을 중지하였다.

모형을 적합 시킨 결과를 보면 AIC와 BIC값을 비교해보았을 때 조건부 분포로 포아송분포를 사용한 Poisson INARCH(1) 모형보다는 ZIP-INARCH(1) 모형이, 음이항분포를 사용한 NBINARCH(1) 모형보다 영-과잉 음이항분포를 사용한 ZINB1-INARCH(1), ZINB2-INARCH(1) 모형이 더 작은 값을 갖는 것을 알 수 있다. 영-과잉 음이항분포에서 값이 0으로 수렴할 때 영-과잉 포아송분포에 가까워진다는 점을 고려하면 ZIP-INARCH(1), ZINB1-INARCH(1), ZINB2-INARCH(1) 세 모형은 내포된(nested) 모형이므로 AIC, BIC값을 이용하여 비교가능하다. 세 모형의 AIC, BIC값을 보면 ZINB2-INARCH(1) 모형이 가장 우수함을 알 수 있다. 조건부 분포로 영-과잉 포아송분포보다는 영-과잉 음이항분포를 이용한 모형이 상대적으로 설명력이 높음을 알 수 있으며 이는 자료에 극단적인 관측값이 존재하기 때문인 것으로 생각된다.

또한 ZIP-INARCH(1) 모형이 INARCH(1) 모형보다 적합한지 가능도비 검정(likelihood ratio statistic)을 통해 확인해 볼 수 있다. 귀무가설은 $H_0 : \omega = 0$으로 두고, 다음과 같은 통계량을 이용한다.

$$\text{LRT} = -2\left[\log L(\theta^*) - \log L(\theta^{**})\right],$$

여기서 $\log L(\theta^*)$는 귀무가설 하에서 구한 로그가능도함수이고, $\log L(\theta^{**})$는 대립가설 하에서 구한 로그가능도함수이다. 귀무가설 하에서 LRT값은 근사적으로 상수 0과 $\chi^2(1)$분포의 50 : 50 mixture 분포를 따르는 것으로 생각할 수 있으며, 이러한 분포에서 상위 α만큼의 확률을 주는 값은 $\chi^2(1)$에서 상위 2α만큼의 확률을 주는 값으로 생각할 수 있다 (Self와 Liang, 1987). 계산된 LRT값은 7.9409로 $\chi^2_{0.98}(1) = 5.4119$와 비교해봤을 때 유의함을 알 수 있으며, 콜레라 자료에 영-과잉 현상이 있다는 것 또한 확인할 수 있다.

References

Ferland, R., Latour, A. and Oraichi, D. (2006). Integer-valued GARCH process, *Journal of Time Series Analysis*, **27**, 923–942.

Fokianos, K. (2011). Some recent progress in count time series, *Statistics*, **45**, 49–58.

Grunwald, G. K., Hyndman, R. J., Tedesco, L. and Tweedie, R. L. (2000). Non-Gaussian conditional linear AR(1) models, *Australian & New Zealand Journal of Statistics*, **42**, 479–495.

Hwang, S. Y. and Basawa, I. V. (2011). Asymptotic optimal inference for multivariate branching-Markov processes via martingale estimating functions and mixed normality, *Journal of Multivariate Analysis*, **102**, 1018–1031.

Puig, P. and Valero, J. (2006). Count data distributions: Some characterizations with applications, *Journal of the American Statistical Society*, **101**, 332–340.

Self, S. G. and Liang, K.-Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions, *Journal of the American Statistical Association*, **82**, 605–610.

Yoon, J. E. and Hwang, S. Y. (2015) Integer-valued GARCH models for count time series: Case study, *Korean Journal of Applied Statistics*, **28**, 115–122.

Zhu, F. (2011). A negative binomial integer-valued GARCH model, *Journal of Time Series Analysis*, **32**, 54–67.

Zhu, F. (2012). Zero-inflated Poisson and negative binomial integer-valued GARCH models, *Journal of Statistical Planning and Inference*, **142**, 826–839.

조건부 포아송 및 음이항 분포를 이용한 영-과잉 INGARCH 자료 분석

윤재은[a] · 황선영[a,1]

[a]숙명여자대학교 통계학과

(2015년 5월 26일 접수, 2015년 6월 8일 수정, 2015년 6월 9일 채택)

요 약

영-과잉(zero-inflation) 현상은 최근 계수(count) 시계열 분석의 주요토픽으로 다루어지고 있다. 본 논문에서는 영-과잉 계수 시계열의 변동성을 연구하고 있다. 기존의 정수형 모형인 INGARCH(integer valued GRACH) 모형에 조건부 포아송 및 조건부 음이항 분포를 사용하여 변동성에 영-과잉 현상을 추가하였다. 모수 추정 방법으로 EM알고리즘을 사용하였으며 국내 콜레라 발생건수에 적용시켜 보았다.

주요용어: 영-과잉 변동성 모형, 정수값-GARCH (INGARCH), 조건부 포아송 및 음이항 분포

[1]교신저자: (140-742) 서울특별시 용산구 청파로47길 100, 숙명여자대학교 통계학과.
E-mail: shwang@sookmyung.ac.kr