

재무 시계열 자료의 이상치 탐지 기법 연구

하명호¹ · 김삼용²

¹중앙대학교 통계학과, ²중앙대학교 통계학과

(2009년 12월 접수, 2010년 1월 채택)

요약

본 연구에서는 재무 시계열 자료를 분석하는데 있어 유용하게 쓰이는 이분산성 시계열 모형하에서 이상치 탐지 기법을 적용하여 그 효율성을 보이고자 한다. 먼저 GARCH 모형과 GARCH 모형하에서 이상치 탐지 기법에 대해 소개하고, 적용된 방법이 기존의 전통적인 이상치 탐지 방법보다 성능이 우수함을 시뮬레이션과 실제 KOSPI 자료에 적합시켜 입증하였다.

주요용어: 이상치, GARCH 모형, KOSPI 자료.

1. 서론

시계열 자료의 분석에서 가장 널리 쓰이는 통계적 기법으로는 Box와 Jenkins (1976)가 제안한 ARIMA (Autoregressive Integrated Moving Average) 모형이 있으며, 이 기법을 이용하여 여러 종류의 시계열 자료들을 분석하고 예측하고 있다. 또한 환율(exchange rates)이나 주가(stock prices)의 변동성을 예측하는데 오차의 분산이 자기 회귀적으로 변하는 이분산성 모형으로 Engle (1982)에 의해 처음 제시된 ARCH 모형이 있다. ARCH 모형의 일반화된 모형으로 Bollerslev (1986)에 의해 제시된 GARCH 모형 등 여러 가지 이분산성 시계열 모형도 재무 시계열 자료를 분석하고 예측하는데 적용되고 있다. 그러나 모형을 적용함에 있어 이상치의 존재는 추정과 예측에 있어 매우 심각한 문제를 일으킬 수 있다. 이러한 문제를 해결하기 위해 많은 연구자들이 다양한 통계적 방법을 적용하여 이상치를 탐지하는 방법을 연구하고 있다. Franses와 Ghijssels (1999)은 GARCH 모형을 이용한 이상치의 탐지 방법을 제안하였고, Charles와 Darné (2006)는 재무 시계열 자료에서 GARCH 모형을 이용한 이상치 탐지 기법의 효율성을 보였다.

본 연구에서는 이러한 최신의 연구동향의 바탕 아래 KOSPI 자료의 이상치를 탐지하는데 있어 GARCH 모형하에서 Charles와 Darné (2006)이 제시한 이상치 탐지 기법의 효율성을 보이고자 한다.

2. 이상치 탐지 기법

기존의 이상치 탐지 기법은 오차의 분산이 동일하다는 가정을 한 이상치 탐지 기법이다. 따라서 오차의 분산이 동일하지 않은 이분산성 시계열 모형에서 이와 같은 이상치 탐지 기법을 사용하는 것은 적절치 않다. 여기서 이분산성 시계열 모형인 GARCH 모형에 대해 살펴보고 Charles와 Darné (2006)가 제안한 GARCH 모형에서의 이상치 탐지 방법에 대해 살펴 보겠다.

이 논문은 2008학년도 정부(교육인적자원부)의 재원으로 한국 대학교육협의회 대학교수 국내교류 연구비 지원에 의한 것임.

²교신저자: (281-207) 서울시 동작구 흑석동 221, 중앙대학교 통계학과, 교수. E-mail: sahm@cau.ac.kr

2.1. GARCH 모형

GARCH 모형은 오차의 분산이 자기 회귀적으로 변하는 조건부 이분산자기회귀모형인 ARCH 모형의 일반화된 모형으로 Bollerslev (1986)에 의해 제시된 모형으로 다음과 같다.

e_t 가 GARCH(p, q) 모형을 따른다고 하면

$$\begin{aligned} e_t &= \sqrt{h_t} \epsilon_t, \quad t = 1, 2, \dots, n, \quad \epsilon_t \sim iid(0, 1) \\ h_t &= \alpha_0 + \sum_{i=1}^p \alpha_i e_{t-i}^2 + \sum_{j=1}^q \beta_j h_{t-j} \end{aligned} \quad (2.1)$$

이다. 여기서 $\alpha_0 > 0$, $\alpha_i > 0$, $\beta_j > 0$, $\sum_{i=1}^{\max(p,q)} (\alpha_i + \beta_j) < 1$ 이다.

본 연구에서는 다음과 같은 GARCH(1,1) 모형을 고려하기로 한다.

$$\begin{aligned} e_t &= \sqrt{h_t} \\ h_t &= \alpha_0 + \alpha_1 e_{t-1}^2 + \beta_1 h_{t-1} \\ \epsilon_t &\sim iid(0, 1). \end{aligned} \quad (2.2)$$

일반적으로 GARCH(1,1) 모형은 재무 자료의 변동성을 ARCH 모형보다 효율적으로 탐지 할 수 있다.

2.2. 이상치 탐지 기법

이상치 검출 방법으로 $\pm 3\sigma$ 를 넘는 값을 이상치로 판명하는 방법과 Charles와 Darné (2006)가 제안한 방법을 사용하였다. $\pm 3\sigma$ 를 넘는 값을 이상치로 판명하는 방법에서 이상치가 발견되면 이상치 값을 전체 평균으로 대체하였다. 또한 Churtes와 Darné (2006)가 제안한 방법에서는 다음과 같은 절차를 따라 통하여 e_t 를 e_t^* 로 대체하였다.

단계 1: 잔차에 대해 다음과 같이 GARCH(1,1) 모형

$$h_t = \alpha_0 + \beta h_{t-1} + \alpha_1 e_{t-1}^2$$

에 대해 α_0 , α_1 , β 를 추정하고,

$$\hat{v}_t = e_t^2 - \hat{h}_t$$

라고 한다. 여기서 e_t 는 관측된 자료이다.

단계 2: $\tau = 1, 2, \dots, n$ 에 대해

$$\hat{\zeta}(\tau) = \frac{\sum_{t=\tau}^n x_t \hat{v}_t}{\sum_{t=\tau}^n x_t^2}$$

를 계산한다. 여기서 x_t 는 다음과 같다.

$$\begin{cases} x_t = 0, & t < \tau, \\ x_t = 1, & t = \tau, \\ x_{t+k} = -\pi_k, & t > \tau. \end{cases}$$

표 3.1. 정확히 이상치를 탐지한 횟수($\hat{\sigma}$ 은 로그 수익률의 표본 표준편차)

이상치 부여 값(W)	이상치 탐지 방법	이상치 탐지 수
W = 3.5	$\pm 3\hat{\sigma}$	1
	제시된 방법	71
W = 5.0	$\pm 3\hat{\sigma}$	12
	제시된 방법	94

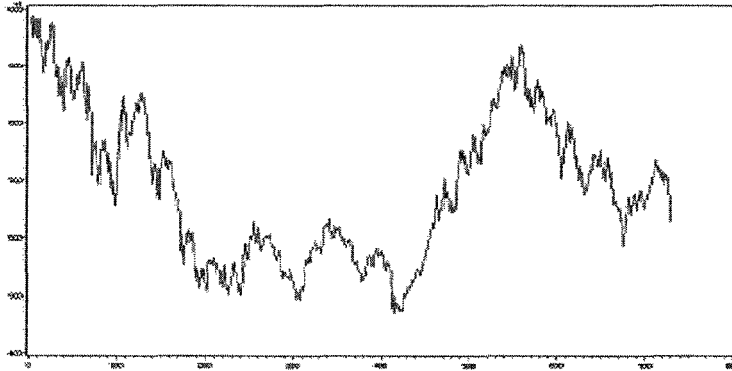


그림 3.1. 원 자료의 시계열 그림

단계 3: $\hat{\tau} = (\hat{\zeta}(\tau)/\hat{\sigma}_a)/(\sum_{t=\tau}^n x_t^2)^2$ 에 대해 $t \max(\hat{\tau})$ 를 계산하고, $t \max(\hat{\tau}) > 10$ 이면, 그 관측치는 이상치이다.

단계 4: $\hat{e}_\tau^* = e_\tau^2 - \hat{\zeta}(\tau)$ 라 하고,

$$e_t^* = \begin{cases} et, & \text{for } t \neq \hat{\tau}, \\ \text{sign}(e_{\hat{\tau}}) \sqrt{e_{\hat{\tau}}^*}, & \text{o/w} \end{cases}$$

라 하자.

단계 5: e_t^* 을 가지고 단계 1부터 $t \max(\hat{\tau})$ 가 10을 넘지 않을 때까지 반복한다.

3. 시뮬레이션 및 실제자료 분석

먼저 시뮬레이션은 GARCH(1,1) 모형에서 모수값 $\phi = 0.3$, $\alpha_0 = 0.1$, $\alpha_1 = 0.1$, $\beta = 0.7$ 을 주고 1,000개의 자료를 생성하였다. 이 중에서 3개의 값을 임의로 선택하여, 이상치를 부여 하였다. 이상치는 3.5와 5이다. 이러한 과정을 각각 100회씩 반복하여 정확히 이상치를 탐지한 횟수를 구하였다.

표 3.1을 보면 이상치 탐지에 있어 $\pm 3\hat{\sigma}$ 를 적용시킨 결과 이상치를 정확히 탐지하는 빈도가 매우 낮았다. 그러나 GARCH 모형하에 이상치를 탐지 한 결과 맞춘 횟수가 매우 높아졌다.

다음으로는 실제자료로서 본 분석에 사용된 자료는 2000년 1월 1일 ~ 2002년 12월 31일까지 총 728개의 KOSPI 자료이다. 우선 이 자료에 대한 시계열 그림은 그림 3.1과 같다. 분석에 있어 정상성을 만족시켜야 하므로 이 자료에 대해 로그 차분하고 100을 곱한 로그 수익률을 사용하였다. 로그 수익률에 대한 그림은 그림 3.2와 같다. 또한 이상치 대체 후의 로그 수익률에 대한 시계열 그림은 그림 3.3과 3.4와 같다. 그림 3.2와 비교해 보면 이상치의 값들이 대체된 것을 볼 수 있다.

이상치 대체 전·후의 추정치를 비교하고 잔차의 정규성 검정을 하겠다. 우선 추정치는 표 3.2와 같다.

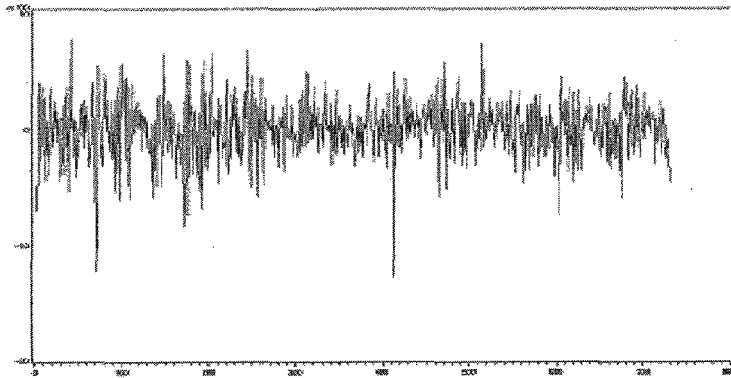


그림 3.2. 로그 수익률($(\log y_t - \log y_{t-1}) \times 100$) 시계열 그림

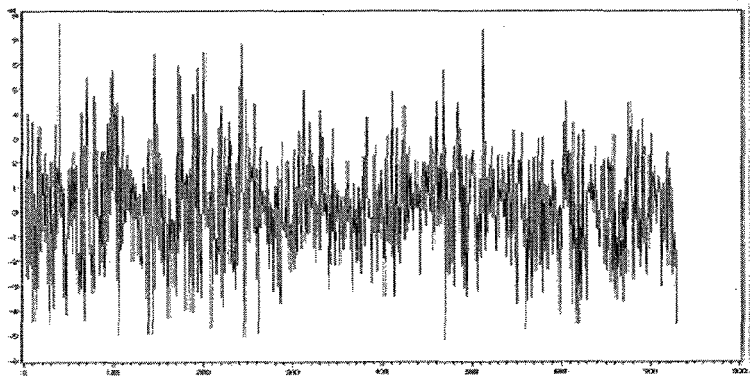


그림 3.3. 이상치 대체 후 시계열 그림(3σ)

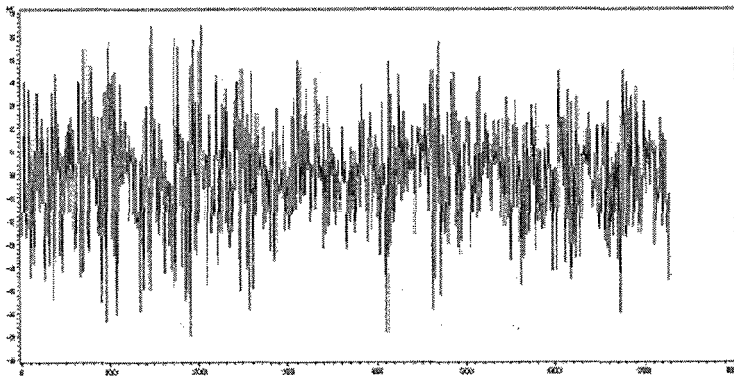


그림 3.4. 이상치 대체 후 시계열 그림(제시된 방법)

추정치를 비교한 결과 3σ 에서는 $\hat{\alpha}_0$ 를 제외한 나머지 추정치가 0에 매우 가깝게 나타나 더 이상 GARCH의 형태가 나타나지 않는 것을 알 수 있다. 그리고, 제시된 방법에서는 이상치 대체 후의 $\hat{\alpha}_0$, $\hat{\beta}$ 의 표준오차가 상당히 작아지는 것을 볼 수 있다. 그러나 $\hat{\alpha}_1$ 의 표준오차는 약간 커졌다.

표 3.2. 이상치 대체 전, 후의 추정치 비교

		추정치	표준 오차
이상치 대체 전	$\hat{\alpha}_0$	0.6883	0.2942
	$\hat{\alpha}_1$	0.0905	0.0274
	$\hat{\beta}$	0.7913	0.0750
$3\hat{\sigma}$	$\hat{\alpha}_0$	4.5369	0.2188
	$\hat{\alpha}_1$	0.0000	0.0000
	$\hat{\beta}$	0.0000	0.0000
제시된 방법	$\hat{\alpha}_0$	0.3284	0.1746
	$\hat{\alpha}_1$	0.0842	0.0297
	$\hat{\beta}$	0.8507	0.0558

표 3.3. $3\hat{\sigma}$, 제시된 방법의 방법에 따른 이상치 대체 전, 후 JB 검정 결과

	이상치 대체 전	$\hat{\mu} \pm 3\hat{\sigma}$	제시된 방법
통계량	202.9245	21.6422	4.7848
p-value	<.0001	<.0001	0.0914

잔차의 정규성 검정 방법으로는 Jarque-Bera 검정을 실시하였다. Jarque-Bera 검정 통계량(JB)은 다음과 같다.

$$JB = \frac{n}{6} \left(S^2 + \frac{(K-3)^2}{4} \right),$$

여기서, $S = \hat{\mu}_3/\hat{\sigma}^3$, $K = \hat{\mu}_4/\hat{\sigma}^4$ 이며, $\hat{\mu}_3$, $\hat{\mu}_4$ 는 각각 3차, 4차 표본 중심 적률이다.

정규성 검정 결과는 표 3.3과 같다.

검정 결과 이상치 대체 전과 $3\hat{\sigma}$ 를 사용한 이상치 검출에서는 잔차가 정규분포를 만족한다고 볼 수 없지만, 이상치 대체 후에는 잔차가 정규분포를 만족한다고 볼 수 있다.

시뮬레이션에서 제시된 방법 방법이 $3\hat{\sigma}$ 를 사용한 방법보다 이상치 탐지 성능이 매우 높았으며, 실제 자료에서는 $3\hat{\sigma}$ 의 잔차는 정규성을 만족한다고 볼 수 없지만, 제시된 방법에 대한 잔차는 정규성을 만족시키는 것으로 나타났다. 이러한 것을 종합해 볼 때 제시된 방법이 $3\hat{\sigma}$ 방법보다 성능이 우수하다 할 수 있다.

4. 결과

본 연구에서는 추정과 예측에 있어 매우 심각한 문제를 일으킬 수 있는 이상치의 탐지 방법에 대한 성능 비교를 위해 GARCH 모형을 소개하고 GARCH 모형하에서 시뮬레이션과 이상치 탐지 횟수와 실제 자료의 잔차 검정을 통하여 기존의 방법과 성능을 비교해 보았다. 향후 연구과제는 Power GARCH 모형, Threshold PARCH 모형과 같은 더욱 정교한 모형을 가지고 이상치 탐지에 대한 성능평가를 실시하는 것이다.

참고문헌

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity, *Journal of Econometrics*, 31, 307-327.

- Box, G. E. P. and Jenkins, G. M. (1976). *Time Series Analysis: Forecasting and Control*, Holden Day, San Francisco.
- Charles, A. and Darné, O. (2006). Outliers and GARCH models in financial data, *Journal of Economics Letters*, **86**, 347-352.
- Engle, R. F. (1982). Autoregressive conditional heteroskedasticity with estimates of the variance of UK inflation, *Econometrica*, **50**, 987-1007.
- Fox, A. J. (1972). Outliers in time series, *Journal of Royal Statistical Society B*, **34**, 350-363.
- Franses, P. H. and Ghijsels, H. (1999). Additive outliers, GARCH and forecasting volatility, *International Journal of Forecasting*, **15**, 1-9.

A Study on Outlier Detection Method for Financial Time Series Data

M.H. Ha¹ · S. Kim²

¹Department of Statistics, Chung-Ang University; ²Department of Statistics, Chung-Ang University

(Received December 2009; accepted January 2010)

Abstract

In this paper, we show the performance evaluation of outlier detection methods based on the GARCH model. We first introduce GARCH model and the methods of outlier detection in the GARCH model. The results of small simulation and the real KOSPI data show the out-performance of the outlier detection method over the traditional method in the GARCH model.

Keywords: Outliers, GARCH model, KOSPI data.

This paper was supported by Korean Council for University Education, grant funded by Korean Government(MOEHRD) for 2008 Domestic Faculty Exchange.

²Corresponding author: Professor, Department of Statistics, Chung-Ang University, 221 Heukseok-Dong, Dongjack-Gu, Seoul 156-756, Korea. E-mail: sahm@cau.ac.kr