Journal of the Korean Data & Information Science Society
2010, **21**(6), 1041–1049

한국데이터정보과학회지

일반화 자기회귀 조건부 이분산 모형을 이용한 한국프로야구 관중수의 예측

이장택[1] · 방소영[2]

[12]단국대학교 정보통계학과

접수 2010년 8월 2일, 수정 2010년 10월 4일, 게재확정 2010년 10월 13일

요 약

한국프로야구에서 관중수는 프로야구 발전을 위한 가장 큰 수입원이며 프로야구팀의 관심사이므로 수요예측 모형이 있다면 프로야구구단들은 관중유치 전략을 세우는데 도움이 될 것이다. 이러한 이유로 본 연구에서는 한국프로야구 관중수를 예측하는 모형을 제안하고자 하며 제한된 여건 속에서 관중수에 영향을 미치는 이용 가능한 대부분의 변수들을 고려하였다. 종속변수는 로그관중수로 두고 다양한 독립변수와 오차항의 분산을 등분산, 조건부 이분산을 가정한 여러 가지 일반화 자기회귀 모형, 오차항의 분포가 t분포를 따른다는 가정을 이용한 일반화 자기회귀 조건부 이분산 모형들을 서로 비교하였는데, 그 결과 고려된 모형 중에서는 t분포를 가정한 일반화 자기회귀 조건부 이분산 모형이 가장 예측력이 뛰어났다.

주요용어: 로그관중수, 수요예측 모형, 일반화 자기회귀 조건부 이분산 모형, 한국프로야구.

1. 서 론

프로야구팀들의 중요한 관심사중의 하나는 관중수인데 관중의 유치는 프로야구 발전을 위한 수입원과 밀접한 관계가 있기 때문이다. 이와 같은 이유로 세계야구를 선도하는 메이저리그를 대상으로 관중수를 예측하는 연구들이 많이 수행되었는데, 승률이 관중수에 중요한 영향을 미친다는 연구 (Fort와 Rosenman, 1999; Meehan 등, 2007), 승률보다는 다른 요인이 관중수에 더 큰 영향을 미친다는 분석 (Gifis와 Sommers, 2006; Chupp 등, 2007), 마이너리그 관중수에 영향을 미치는 요인에 대한 연구 (Paul 등, 2009), GARCH 모형을 이용한 팀의 승리와 메이저리그 관중수에 관한 연구 (Michael, 2009) 등이 있으며, 한국프로야구에 관한 관중수를 예측하는 연구로는 Lee (2006)가 있다.

관중수에 영향을 미치는 요인에는 날씨, 경기시간, 홈구장의 시설, 이벤트, 구단의 승률 등과 같은 많은 변수들이 있겠지만 현실적으로 이런 변수들을 모두 고려하기는 매우 어렵다. 따라서 본 연구에서는 선행연구들을 참고로 하여 한국프로야구에서 영향을 많이 줄 것으로 간주되는 관측 가능한 여러 가지 주요 변수들을 고려하여 한국프로야구 관중수를 예측하는 모형을 제안하고자 한다. 일반적으로 관중수는 복합적인 요인들에 의해 변동성이 큰데 이러한 변동성을 등분산 모형으로 설명하는 것은 한계가 있다고 생각되어 조건부 이분산성을 고려한 모형 중 일반적으로 많이 사용되는 일반화 자기회귀 조건부 이분산 모형 (GARCH)을 활용하였다. GARCH 모형에 대한 최근 연구들은 Park과 Lee (2007), Lee (2007), Lee와 Ha (2007), Lee (2009) 등이 있다.

[1] 교신저자: (448-701) 경기도 용인시 수지구 죽전동 126번지, 단국대학교 정보통계학과, 교수.
E-mail: jtlee@dankook.ac.kr
[2] (448-701) 경기도 용인시 수지구 죽전동 126번지, 단국대학교 정보통계학과, 석사과정.

본 논문의 구성은 다음과 같다. 2절에서는 데이터의 구성, 데이터의 처리, 통계분석을 위한 소프트웨어 및 모형평가 기준에 대하여 언급하였으며, 3절에서는 고려된 세 가지 모형에 대한 설명과 결과분석을 하고, 검증용 데이터를 이용한 예측력 분석 결과를 설명하였다. 끝으로 4절에서는 연구결과의 요약 및 그 의의에 대해 언급하였다.

2. 자료 분석 및 모형평가

2.1. 자료의 구성

표 2.1 사용된 변수들에 대한 설명

변수 종류	변수 이름	설명
종속변수	로그관중수	자연로그를 취한 홈구장의 관중수
독립변수 (시간관련)	공휴일	해당일에 속하면 1, 속하지 않으면 0
	주말 (토,일)	(공휴일과 주말에 속하지 않는 경우 평일)
	4월	해당 월에 속하면 1, 속하지 않으면 0
	5월	
	6월	
	7월	
	8월	
	9월	
	더블헤더	경기가 더블헤더인 경우 1, 아니면 0
독립변수 (구단관련)	플레이오프	직전년도 플레이오프에 진출했던 구단이면 1, 진출 못한 구단이면 0
	승률	직전경기까지의 누적 승률
	삼성	해당 구단에 속하면 1, 속하지 않으면 0
	현대	
	기아	
	두산	
	SK	
	롯데	
	LG	
독립변수 (자기상관관련)	로그관중수$_{t-1}$	1시차 전의 로그 관중수
	로그관중수$_{t-2}$	2시차 전의 로그 관중수

본 연구에서는 2000년부터 2006년 동안의‘한국프로야구 연감’에 기록된 경기와 관련된 여러 자료를 이용하였다. 총 3,668경기를 통해 2000년부터 2005년도의 자료를 훈련자료로, 2006년도의 자료를 검증자료로 사용하였다. 종속변수는 로그변환을 한 관중수 (로그관중수)이고 독립변수는 크게 시간, 구단, 자기상관에 관한 3가지 요인으로 나누어 분석하였으며 변수에 관한 자세한 설명은 표 2.1에 제시되어 있다. 그리고 공휴일에는 식목일, 제헌절, 어린이날, 석가탄신일, 현충일, 광복절, 개천절을 포함시켰다.

2.2. 데이터처리 및 통계분석

통계패키지 SPSS 17K와 MATLAB을 이용하여 회귀방정식을 구하고 잔차의 정규성검정, 단위근 검정 및 ARCH 검정을 시행하였다.

2.3. 모형평가 기준

본 연구에서는 제안된 모형들의 효율성을 서로 비교하기 위하여 일반적으로 많이 사용되는 추정량 선택기준인 평균제곱오차의 제곱근 (RMSE, root mean square error)과 평균절대편차 (MAD, mean absolute deviation)를 사용하였다. 이 경우 시간 i의 관중수를 y_i, 추정된 관중수를 $\hat{y}_i$, 총 게임수를 n이라고 두면, RMSE와 MAD는 각각 다음 식과 같이 정의되며, 추정량들의 RMSE와 MAD를 계산하여 값이 가장 작은 것이 제일 좋은 추정량이라 할 수 있다.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \widehat{y_i})^2}{n}} \quad , \quad MAD = \frac{\sum_{i=1}^{n}|y_i - \widehat{y_i}|}{n}$$

3. 모형 설정

3.1. 모형설정 및 결과분석

시계열 변수를 이용해서 회귀분석을 시도하는 경우, 단위근의 유무를 확인하는 이유는 허구적 회귀를 피하기 위해서이다. 허구적 회귀 (spurious regression)란 단위근이 있는 시계열은 일시적으로 확률적 추세 (선형적 추세)가 있는 것처럼 나타날 수 있는데, 이로 인해 무의미한 회귀계수가 유의하게 추정되는 현상이다 (김명직 등, 2003). 따라서 회귀분석 이전에 시계열 자료에 대한 단위근 검정을 수행하여 시계열의 확률적 추세 존재유무를 파악할 필요가 있다. 본 연구에서는 로그관중수에 대하여 최근 많이 사용되고 있는 단위근 검정법인 ADF검정 (Augmented Dickey Fuller test)를 사용하여 분석하였다.

ADF검정은 오차항이 자기상관 (autocorrelated)되어 있는 경우에 있어서 자기상관의 효과를 완화시키기 위해 오차항이 백색오차 (white noise)가 되도록 충분한 수의 시차 차분항 (lagged difference terms)을 추가하는데, 차분추가항의 차수는 표본자료의 정보량을 잘 반영하는 Akaike의 정보기준 AIC에 의거하여 AIC값이 최소가 되는 모형을 선택하였다. 표 3.1의ADF의 단위근 검정 결과를 보면 2차 차분한 경우에 추세항을 모두 포함한 검정식을 사용하여 분석한 결과, 유의수준 1%에서 단위근이 존재한다는 귀무가설을 기각할 수 있어서 시계열에 확률적 추세가 없고 따라서 이로 인한 허구적 회귀문제를 피할 수 있다고 할 수 있다.

표 3.1 ADF의 단위근 검정

구 분	ADF검정		
	시차	ADF 통계치	유의확률
로그관중수	2	-44.8498	0.000

본 연구에서 사용되는 기본적인 통계모형은 다음 <모형1>과 같다.

< 모형1 >

$$\begin{aligned}
\text{로그관중수}_t &= \alpha_0 + \alpha_1\text{로그관중수}_{t-1} + \alpha_2\text{로그관중수}_{t-2} + \alpha_3\text{더블헤더}_t + \alpha_4\text{공휴일}_t + \alpha_5\text{주말}_t \\
&\quad + \alpha_6 4\text{월}_t + \alpha_7 5\text{월}_t + \alpha_8 6\text{월}_t + \alpha_9 7\text{월}_t + \alpha_{10} 8\text{월}_t + \alpha_{11} 9\text{월}_t + \alpha_{12}\text{플레이오프}_t + \alpha_{13}\text{승률}_t \\
&\quad + \alpha_{14}\text{삼성}_t + \alpha_{15}\text{현대}_t + \alpha_{16}\text{기아}_t + \alpha_{17}\text{두산}_t + \alpha_{18}\text{SK}_t + \alpha_{19}\text{롯데}_t + \alpha_{20}\text{LG}_t + \epsilon_t \\
&= \boldsymbol{x_t}'\boldsymbol{\alpha} + \alpha_1\text{로그관중수}_{t-1} + \alpha_2\text{로그관중수}_{t-2} + \epsilon_t
\end{aligned}$$

<모형 1>의 회귀계수를 추정한 표 3.2에서 VIF와 R^2값에 대해 살펴보면 VIF가 전부 10을 넘지 않으므로 다중공선성 문제는 없다고 할 수 있으며, 모형의 설명력 측도인 R^2값은 독립변수들이 종속변수의 전체 분산의 65.6%를 설명한다고 할 수 있다. 또한 t-검정을 통한 개별 회귀계수들은 변수 기아를 제외하고 유의수준 5%에서 전부 유의한 것으로 나타나서 고려된 대부분의 독립변수들이 관중수에 영향을 미친다고 판단할 수 있다.

3.2. 조건부 이분산성을 고려한 추가 모형 설정

<모형1>에 대한 잔차들의 왜도는 -0.401, 첨도는 1.119로 첨도의 값이 0보다 커서 정규분포보다 꼬리가 두터운 분포를 갖는다. 이 경우 변수 로그관중수 대신 관중수를 사용하면 첨도의 값은 5.14로서 꼬리가 매우 두터운 분포임을 알 수 있다. 또한 잔차가 정규분포의 형태를 갖는지를 확인하기 위해 Jarque-Bera 검정을 실시하였는데, 유의확률이 0.001로 유의수준 1%에서 정규분포를 따르지 않는 것으로 나타났으며, Engle (1982)의 자기회귀 조건부 이분산 (Autoregressive Conditional Heteroskedasticity: ARCH) 모형의 도입여부를 위하여 Engle (1982)의 ARCH 검정을 수행하면 ARCH 효과 존재 여부를 파악 할 수 있는데, 표 3.3을 통하여 유의수준 1%에서 ARCH 효과가 있다고 할 수 있으므로 오차항을 꼬리가 두터운 t분포로 가정한 여러 가지 ARCH류 모형을 고려하여 분석하는 것이 타당하다고 간주되어진다.

표 3.2 <모형 1>에 대한 회귀계수의 추정

	모형1		
	계수	유의확률	VIF
(상수)	2.374	0.000	-
더블헤더	-0.343	0.000	1.077
공휴일	0.648	0.000	1.045
주말	0.499	0.000	1.032
4월	0.510	0.000	4.966
5월	0.578	0.000	6.088
6월	0.421	0.000	5.597
7월	0.471	0.000	4.759
8월	0.483	0.000	5.365
9월	0.327	0.000	4.707
플레이오프	0.073	0.004	1.514
승률	1.148	0.000	1.694
삼성	0.115	0.014	2.295
현대	-0.279	0.000	2.160
기아	-0.020	0.624	1.790
두산	0.405	0.000	2.050
SK	0.221	0.000	1.829
롯데	0.270	0.000	1.919
LG	0.648	0.000	2.272
$t-1$	0.351	0.000	2.503
$t-2$	0.184	0.000	2.333
R^2		0.656	

표 3.3 Engle의 ARCH 검정

ARCH 검정 통계치	유의확률
53.5728	0.0000

한편 본 연구에서는 여러 가지 ARCH 모형을 고려하는 것도 좋지만 Bollerslev (1986)의 GARCH (Generalized ARCH) 모형 중 단순한 GARCH(1,1)모형으로도 반복적 대입과정을 통하여 ARCH(∞) 모형으로 다시 쓸 수 있기 때문에 ARCH 모형대신 GARCH(1,1) $-t$와 Nelson (1991)의 EGARCH (Exponential GARCH) 중 간단한 EGARCH(1,1) $-t$의 두 가지 모형을 고려하였다. <모형 2>와 <모

형 3>에서 벡터 $\boldsymbol{x}_t$는 <모형 1>의 정의와 같으며, $\boldsymbol{\beta}$와 $\boldsymbol{\gamma}$는 각각 $\boldsymbol{x}_t$에 해당되는 회귀계수벡터이다.

$$< \text{모형}2 > AR(2)-GARCH(1,1)-t$$

$$\text{로그관중수}_t = \boldsymbol{x}_t{}'\boldsymbol{\beta} + \beta_1 \text{로그관중수}_{t-1} + \beta_2 \text{로그관중수}_{t-2} + \epsilon_t, \epsilon_t = e_t\sigma_t, \quad e_t \sim iid(0,1)$$

$$\sigma_t^2 = \omega + \phi\epsilon_{t-1}^2 + \psi\sigma_{t-1}^2$$

$$< \text{모형}3 > AR(2)-EGARCH(1,1)-t$$

$$\text{로그관중수}_t = \boldsymbol{x}_t{}'\boldsymbol{\gamma} + \gamma_1 \text{로그관중수}_{t-1} + \gamma_2 \text{로그관중수}_{t-2} + \epsilon_t, \epsilon_t = e_t\sigma_t, \quad e_t \sim iid(0,1)$$

$$\log(\sigma_t^2) = \omega + \psi\log(\sigma_{t-1}^2) + \phi\left[\frac{|\epsilon_{t-1}|}{\sigma_{t-1}} - E\left(\frac{|\epsilon_{t-1}|}{\sigma_{t-1}}\right)\right] + L\frac{\epsilon_{t-1}}{\sigma_{t-1}}$$

표 3.4 GARCH(1,1)$-t$ 모형과 EGARCH(1,1)$-t$ 모형의 계수 추정

	모형 2			모형 3		
	계수	표준오차	t	계수	표준오차	t
(상수)	2.2659	0.1088	20.8268	2.2907	0.10883	21.0477
더블헤더	-0.33002	0.026417	-12.4927	-0.327	0.026386	-12.3928
공휴일	0.62043	0.065386	9.4887	0.60285	0.062943	9.5778
토일	0.48775	0.020499	23.7941	0.4851	0.020345	23.8441
4월	0.55574	0.058204	9.5481	0.56183	0.058706	9.5702
5월	0.61253	0.056569	10.8281	0.61245	0.056795	10.7836
6월	0.48673	0.056031	8.6869	0.48933	0.056428	8.6717
7월	0.49004	0.057056	8.5887	0.49225	0.057232	8.6009
8월	0.49937	0.056131	8.8966	0.50193	0.05651	8.8821
9월	0.36173	0.055795	6.4832	0.35421	0.056142	6.3091
플레이오프	0.06956	0.022818	3.0484	0.066945	0.022656	2.9548
승률	1.1972	0.096862	12.3596	1.198	0.096103	12.4662
삼성	0.10839	0.044742	2.4226	0.116	0.044065	2.6324
현대	-0.27226	0.042161	-6.4577	-0.27358	0.04168	-6.564
기아	-0.02154	0.038318	-0.562	-0.01614	0.03797	-0.4251
두산	0.40446	0.039967	10.1198	0.4069	0.039559	10.2859
SK	0.20128	0.039491	5.097	0.20102	0.039021	5.1516
롯데	0.31215	0.03889	8.0266	0.32329	0.038555	8.385
LG	0.6428	0.043145	14.8987	0.65165	0.042636	15.2841
로그관중수$_{t-1}$	0.37185	0.016044	23.1767	0.3724	0.015925	23.3848
로그관중수$_{t-2}$	0.17223	0.014962	11.5106	0.16846	0.014895	11.3097
ω(상수)	0.073757	0.016621	4.4377	-0.24787	0.057005	-4.3482
ψ	0.63374	0.064795	9.7807	0.79034	0.047576	16.6123
ϕ	0.13904	0.025366	5.4814	0.27083	0.040043	6.7634
L	-	-	-	-0.01058	0.019639	-0.5387
자유도	8.0646	1.22	6.6103	8.2281	1.2797	6.4295

모형의 추정은 Matlab에서 제공하는 GARCHSET와 GARCHFIT 절차를 이용하였으며 그 결과인 표 3.4를 보면 t분포의 자유도는 8.0646로 낮게 추정되었는데, 보통 자유도가 20보다 작은 경우에는 두꺼운 꼬리를 갖는 분포로 볼 수 있으므로 잔차들의 분포가 정규분포보다 좀 더 두터운 꼬리를 갖는다고 할 수 있으며 따라서 오차항을 t분포로 가정하는 것이 타당하다는 것을 알 수 있다. 그리고 지속성 모수 $\phi+\psi$은 값이 1에 가까울수록 현재의 변동성 수준이 유사한 수준으로 장래에도 지속될 가능성이 높다고 할 수 있으므로 표 3.4의 <모형 2>를 보면 그 값이 0.77278로 1에 비교적 가깝기 때문에 현재와 유사한 변동성 수준이 장래에도 지속될 가능성이 높은 것으로 보인다. 레버리지 효과란 음의 충격이 동

일한 크기의 양의 충격보다 다음 기 변동성에 더 큰 영향을 준다는 것인데, <모형 3>을 보면 레버리지 L은 t통계량의 값으로부터 유의수준 5%에서 통계적으로 유의하지 않다고 해석할 수 있다. 이러한 이유의 한 가지 가능성은 고정 팬의 영향으로 설명할 수 있는데, 만약 구단의 성적이 좋지 않아 관중수가 줄어들 것이라고 기대되어도 꾸준히 홈구장을 방문하는 충성도가 높은 고정 팬들이 존재하므로 비대칭 효과가 적다고 할 수 있으며 이 점은 대부분의 금융데이터들과 다른 점이라고 볼 수 있다.

3.3. 예측력 분석

본 절에서는 앞에서 설정한 ARCH타입의 모형을 토대로 <모형 1>부터 <모형 3>까지의 여러 가지 모형들을 비교하고자 한다. 2000년부터 2005년까지 6년 동안의 3,164개의 경기 자료를 사용하여 만든 모형을 이용하여 2006년 한 해 동안의 504경기의 관중수를 예측하였는데, 표 3.5는 3가지 모형으로 2006년 관중수의 실제값과 예측값을 이용하여 구한 RMSE와 MAD를 제시한 결과이다. RMSE와 MAD를 보면 예상했던 바와 같이 GARCH모형을 사용하여 만든 <모형 2>가 가장 좋은 모형으로 나타나나 EGRACH를 사용한 <모형 3>과는 큰 차이가 없다. 그림 3.1은 RMSE와 MAD가 가장 작은 <모형 2>를 이용하여 2006년도 관중수 및 예측값을 나타낸 도표이다. 전반적으로 현대가 가장 예측을 잘하는 것으로 보이며, 두산, 롯데, LG도 추세는 잘 예측하지만 갑자기 많아지는 관중수를 추적하여 예측하는 것은 무리가 있다고 할 수 있는데, 아마도 세 팀은 타 팀에 비하여 관중수도 많고 구장이 넓기 때문에 관중수의 변동이 심하다고 할 수 있겠다.

표 3.5 세 가지 모형에 대한 RMSE와 MAD

팀명	RMSE			MAD		
	모형 1	모형 2	모형 3	모형 1	모형 2	모형 3
삼성	1,653.08	1,687.16	1,689.72	1,246.96	1,292.09	1,296.73
한화	2,551.95	2,570.40	2,569.10	1,731.32	1,772.63	1,767.14
현대	881.68	874.59	874.24	659.72	658.59	659.83
기아	2,193.34	2,180.66	2,174.03	1,326.36	1,327.95	1,324.40
두산	6,683.98	6,576.87	6,607.90	4,883.19	4,787.56	4,812.26
SK	3,027.90	3,055.26	3,052.57	1,835.43	1,862.21	1,864.66
롯데	4,260.52	4,023.78	3,993.95	3,140.84	3,008.20	2,989.79
LG	5,941.91	5,887.72	5,896.50	4,283.62	4,305.19	4,316.49
평균	3,399.29	3,357.06	3,357.25	2,388.43	2,376.81	2,378.91
순위	3	1	2	3	1	2

4. 결 론

프로야구 관중수는 구단의 수입과 직결되어 있으므로, 프로야구 구단의 큰 관심사가 된다. 이런 연유로 관중수를 예측하는 모형이 많이 존재하지만, 대부분의 연구는 미국 메이저리그의 관중수를 예측하는 모형에 국한되어있다. 따라서 본 논문은 한국프로야구 관중수를 예측하는 모형을 제시하는데 그 의의를 두고 있다. 본 연구에서는 최종 모형으로 AR(2)-GARCH(1,1)$-t$ 모형을 선택하였는데, 모형분석 결과 로그변환한 관중수 및 ARCH류의 모형으로 분석하는 것이 좀 더 높은 설명력을 가지며, ARCH류의 모형을 설정할 때, 오차항의 분포가 정규분포보다 t분포를 가정한 모형이 더 적합하였다. 또한 지속성 모수의 추정값이 1에 비교적 근접한 수치로 볼 수 있어서 현재와 유사한 변동성 수준이 장래에도 지속될 가능성이 높은 것으로 보이며 EGARCH (1,1)$-t$ 모형보다는 GARCH(1,1)$-t$가 더 적합한 모형으로 선택되었지만 그 차이는 근소했다. 아울러 AR(2)-GARCH(1,1)$-t$ 모형에 대한 MAD의 평균값으로부터 2006년의 예측치가 실제보다 구단별로 평균적으로 2,000명 이상 차이가 난다고 해석될 수 있는데 그

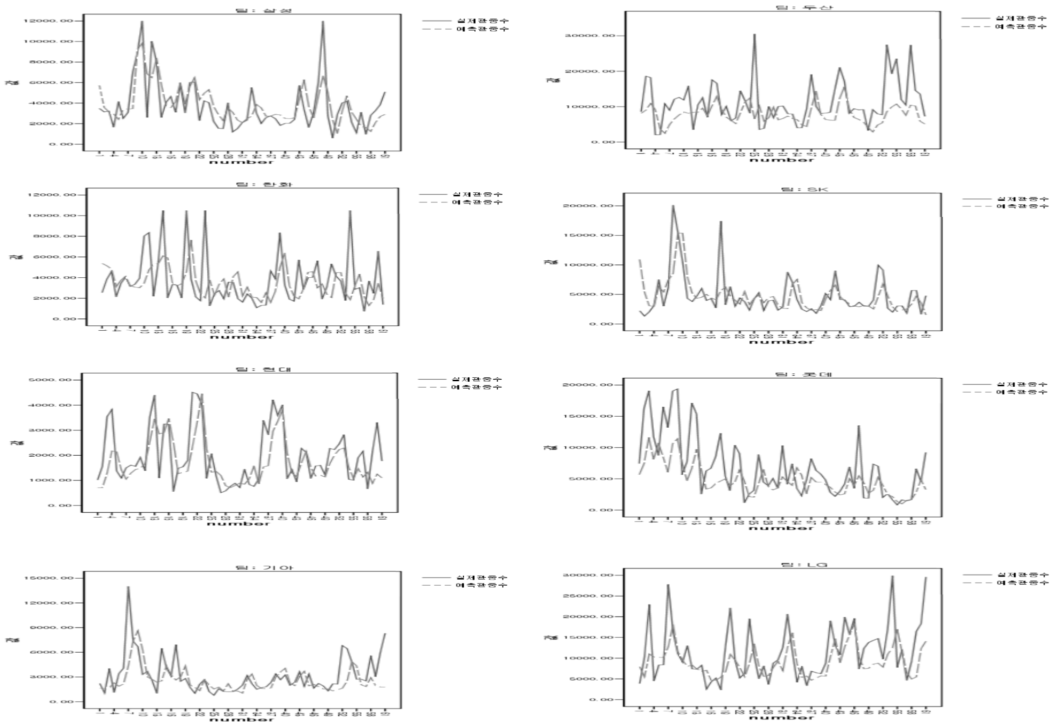

그림 3.1 GARCH$-t$ 모형의 예측값과 실제 관중수

이유는 현실적인 제약여건 관계로 더 많은 요인들을 고려하지 못했기 때문이라고 볼 수 있어 날씨 등의 변수를 추가로 보완한다면 예측력을 좀 더 향상시킬 수 있을 것으로 기대된다.

참고문헌

김명직, 장국현 (2003). <금융시계열분석>, 경문사, 서울.

한국야구위원회 (2000-2006). <한국프로야구 연감>, 한국야구위원회, 서울.

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

Chupp, A., Stephenson, F. and Taylor, R. (2007). Stadium alcohol availability and baseball attendance: Evidence from a natural experiment. *International Journal of Sport Finance*, **2**, 36-44.

Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica*, **50**, 987-1007.

Fort, R. and Rosenman, R. (1999). Streak management. In J. Fizel, E. Gustafson, and L. Hadley (Eds.), *Sports Economics: Current Research*, 119-133.

Gifis, L. S. and Sommers, P. M. (2006). Promotions and attendance in minor league baseball. *Atlantic Economic Journal*, **34**, 513-514.

Lee, O. S. (2007). Sufficient conditions for stationarity of smooth transition ARMA/GARCH models. *Journal of the Korean Data & Information Science Society*, **18**, 237-245.

Lee, T. W. (2009). Numerical study on Jarque-Bera normality test for innovations of ARMA-GARCH models. *Journal of the Korean Data & Information Science Society*, **20**, 453-458.

Lee, T. W. and Ha, J. C. (2007). Testing the domestic financial data for the normality of the innovation based on the GARCH(1,1) model. *Journal of the Korean Data & Information Science Society*, **18**, 809-815.

Lee, Y. H. (2006). The decline of attendance in the Korean professional baseball league: The major league effects. *Journal of Sports Economics*, **7**, 187-200.

Meehan, J. W. Jr., Nelson, R. A. and Richardson, T. V. (2007). Competitive balance and game attendance in major league baseball. *Journal of Sports Economics*, **8**, 563-580.

Michael, C. D. (2009). Analyzing the relationship between team success and MLB attendance with GARCH effects. *Journal of Sports Economics*, **10**, 44-58.

Nelson, D. B. (1991). Conditional heteroscedasticity in asset returns: A new approach. *Econometrica*, **59**, 347-370.

Park, S. Y. and Lee, S. Y. (2007). Modelling KOSPI200 data based on GARCH(1,1) parameter change test. *Journal of the Korean Data & Information Science Society*, **18**, 11-16.

Paul, J. R., Toma, M. and Weinbach, A. P. (2009). The minor league experience: What drives attendance at south atlantic league baseball games? *The Coastal Business Journal*, **8**, 70-84.

Journal of the Korean Data & Information Science Society
2010, **21**(6), 1041–1049

한국데이터정보과학회지

Forecasting attendance in the Korean professional baseball league using GARCH models

Jangtaek Lee[1] · Soyoung Bang[2]

[12]Department of Statistics, Dankook University

Received 2 August 2010, revised 4 October 2010, accepted 13 October 2010

Abstract

In Korean professional baseball, attendance is the largest source of revenue for development of professional baseball and the highest concern of professional baseball teams. So, if there is demand forecasting model, it will be helpful for pennant chasers to work out the strategies for drawing attendance. For this reason, this research intends to suggest the model which estimates Korean professional baseball's attendance and uses all usable variables which have an effect on attendance in limited circumstances. We supposed that dependent variable is attendance as well as several independent variables and error term are homoscedastic variance. And then, we compared the models which assume conditional heteroscedastic variance like GARCH and EGARCH with GARCH$-t$ models which use the assumption that error term's distribution follows student-t distribution. In result of that, we could confirm that the models which were made by using GARCH(1,1)$-t$ made estimates the most accurately among the several models considered.

Keywords: Attendance, demand forecasting model, EGARCH, GARCH, Korean professional baseball.

[1] Corresponding author: Professor, Department of Statistics, Dankook University, Yongin 448-701, Korea. E-mail: jtlee@dankook.ac.kr
[2] Graduate student, Department of Statistics, Dankook University, Yongin 448-701, Korea.