Article

Journal of Sports Economics
2018, Vol. 19(4) 583-598

DOI: 10.1177/1527002516672061
journals.sagepub.com/home/jse
SAGE

Common Factors in Major League Baseball Game Attendance

Young H. Lee[1]

Abstract

This article applies a panel data model with observed common factors to Major League Baseball (MLB) data from 1904 to 2012 to analyze attendance. In particular, it aims to identify common factors. The empirical results suggest that MLB fan preferences were simple in the early years (1904-1957) with respect to common factors and then became multifaceted in later years (1958-2012), because the number of significant common factors increased from four to seven. Time trends and per capita gross domestic product were significant over the whole sample period, but outcome uncertainties and offensive performance, such as slugging performance, became newly significant common factors influencing attendance in later years. This indicates that fans consider not only their home team's characteristics but also the characteristics of the away teams; then, in the modern era, it became critical for the league to implement elaborate business measures to promote competitive balance and slugging performance.

Keywords

attendance, outcome uncertainty, common factors, factor loading, panel data, competitive balance

[1] Department of Economics, Sogang University, Seoul, South Korea

Corresponding Author:
Young H. Lee, Department of Economics, Sogang University, #1 Sinsoo-dong, Mapo-gu, Seoul 136-792, South Korea.
Email: yhnlee@sogang.ac.kr

Introduction

Empirical works on attendance have analyzed diverse issues (uncertainty of outcome hypothesis, stadium effects, fan loyalty, habitual attendance) and now form an extensive literature. Among them, some empirical results have shown that attendance is influenced by the offensive and/or defensive characteristics of visiting teams, even after controlling for the strength of the visiting teams, affecting the uncertainty of game outcomes (e.g., see Buraimo & Simmons, 2008; Coates & Humphreys, 2012; Coates, Humphreys, & Zhou, 2014). Coates, Humphreys, and Zhou used a daily game attendance sample during the 2005-2010 seasons in Major League Baseball (MLB), and they presented the empirical evidence that attendance was influenced by not only the characteristics of the home team but also those of the visiting team.

Recently, Ahn and Lee (2014) applied a factor model to a panel data sample with long-term series in their attendance study. Under the condition that attendance followed a factor structure, there were two contributions. First, the factor model controlled for cross-sectional dependence as well as the omitted variable problem. In fact, they found multiple factors in MLB attendance, and their results indicated that the conventional panel data model may suffer from the omitted variable problem, causing biased estimates. Second, they used team-level panel data from more than 100 seasons to investigate MLB attendance and then examined changes in the determinants of MLB attendance over time. This study is the first application of the factor model in sports economics literature, but factor models have been used extensively in various economic studies. For example, Keloharju, Linnainmaa, and Nyberg (2016) analyzed stock returns and found that the common seasonalities accounted for more than 80% of the variation in individual stock returns.

In this study, we attempted to identify common factors in MLB attendance using a sequential test method. The identification of common factors is important not only academically but also practically. The estimates of attendance regression using identified factors should be more efficient, at least theoretically, than those using all potential factors in set F. Fan demand is influenced not only by team-specific characteristics, such as home team performance, but also by aggregate and league-wide factors. Identification of such factors will help teams and leagues understand fan demand and set policies. In the MLB panel data sample of 1904-1957, the number of common factors was found to be four, whereas it was seven in the 1958-2012 sample. This indicates that fan demands became more multifaceted during the modern era. Attendance is now influenced by more macro or league-wide variables. If the common factors during 1904-1957 remained significant during 1958-2012, then three additional factors were added. However, if any of the four factors during the former period were not significant during the latter period, structural changes may have occurred. Certainly, identification of the true common factors is beneficial.

We describe a sequential test method useful for identification of the factors involved. This method is based on the fact that the rank of the factor loading matrix

decreases by one when a true common factor is deleted from the potential factor set and vice versa. Suppose the common factor matrix F includes q potential factors, and p ($p \leq q$) is the true number of factors. In the case of $p = q$, all variables in F are common factors. Generally, $p < q$, and thus we need to identify p variables in F as true factors. In the first step of the method, each potential factor variable is deleted from F, and the number of factors is estimated. If the estimate is $p - 1$, then the deleted variable is considered a true factor. If the identified factors in the first step are less than p, we conduct the second step as follows: we add each potential factor not identified as a true factor in the first step to the identified factor set and estimate the number of factors. If the number increases by one, then the added variable is considered a true factor. The same steps are repeated until p factors are identified.

The rest of this article is organized as follows: The second section discusses the econometric model and the attendance regression equation. The third section discusses the data set and estimation results. Some concluding remarks are provided in the fourth section.

Econometric Models and Attendance Regression Specification

Ahn and Lee (2014) used team-level panel data for more than 100 seasons to study MLB attendance applying a factor model. In particular, they assumed that team-specific missing variables, such as ticket prices, followed a factor structure. That is, ticket prices of individual teams were determined by a time-varying factor that is common to all teams and the factor loadings corresponding to a common factor are team variant. This assumption may be reasonable if teams consider general economic conditions, as reflected, for example, by the gross domestic product (GDP), and then look at local market characteristics and team performances when determining ticket prices. In this case, general economic conditions is a common factor and local characteristics form factor loadings. Instead of estimating unobserved common factors directly, they assumed that unobservable common factors were linear combinations of observable aggregate variables and then they could estimate the MLB attendance equation with a panel data model with observed common factors. This multifactor specification resolved the omitted variable problem due to missing variables and allowed the use of panel data with a long time series. The number of common factors should be an important concern in the regression model specification. It implies that conventional panel data models with individual effects and/or time effects may suffer from the omitted variable problem and produce biased estimates if omitted common factors are correlated with regressors. For example, a panel data model with only time effects assumes one common factor and that factor loadings are identical across individual teams; a model with only individual effects also assumes one time-invariant common factor but that factor loadings are different across individual teams. Both cases belong to a one-factor model structure with some restrictions, and thus may suffer from the omitted variable problem if there are more than one common factor and the missing common factors are correlated with regressors.

The regression equation in the factor model with observable "proxy" (or "potential") common factors is as follows:

$$y_i = X_i\alpha + \theta_i 1_T + F\beta_i + u_i, \tag{1}$$

for team i, where $y_i = (y_{i1}, y_{i2}, \ldots, y_{iT})'$, $X_i = (x_{i1}, x_{i2}, \ldots, x_{iT})'$, 1_T is a $T \times 1$ vector of ones, $F = (f_1, \ldots, f_T)'$, and u_i is defined similarly to y_i. Here, $f_t = (f_{1t}, f_{2t}, \ldots, f_{qt})'$ is the q vector of factors common to all individual teams; $\beta_i = (\beta_{1i}, \beta_{2i}, \ldots, \beta_{qi})'$ is the $q \times 1$ vector of factor loadings that vary across individual teams. Pooling the models for all teams yields

$$Y = X(I_N \otimes \alpha) + 1_T\theta' + F\mathrm{B} + U, \tag{2}$$

where $Y = (y_1, y_2, \ldots, y_N)$, $X = (X_1, \ldots, X_N)$, $\theta = (\theta_1, \ldots, \theta_N)'$, $\mathrm{B} = (\beta_1, \ldots, \beta_N)$, and U are defined similarly to Y. We assume that the regressor x_{it} and the factors f_t are strictly exogenous; that is,

$$E(u_{it}|x_{i1}, \ldots, x_{iT}, f_1, \ldots, f_T) = 0, \tag{3}$$

for all i. This model structure nests various panel data models that have been used in previous empirical studies on attendance. For example, $q = 0$ implies the standard model with individual effects. The restriction of $q = 1$ with homogenous factor loadings reduces model (1) to the panel model with both individual and time effects in an additive form. The multifactor models with unobserved common factors (Bai, 2009; Ahn, Lee, & Schmidt, 2013) are also available. However, they require large numbers of cross-sectional observations and our sample does not satisfy this condition because there are only 30 teams in MLB. Thus, we put various aggregate variables of potential common factors into the matrix F and assume that all true factors are included in the matrix F.

In the regression Equation 2, we need to estimate the parameters α, θ, and β. They can be estimated by the ordinary least squared (OLS) method, and the OLS estimate of α is consistent and asymptotically normal when T is large. Note that Equation 2 is not an ordinary form. For example, Y is a $T \times N$ dimensional matrix instead of an NT dimensional vector. Thus, the OLS estimates are obtained after the regression equation is vectorized. We estimate the number of factors using the modified Bayesian information criterion (MBIC) estimator, developed by Ahn, Horenstein, and Wang (2013; see Ahn & Lee, 2014, for more details on the estimation process).

The MBIC estimator should decrease by one if a true factor variable is eliminated from the matrix F. Thus, a sequential method can be applied to identify p true common factors from the q potential variables. All f_t are identified as true factors when $p = q$. Generally, $p < q$. The first step in the sequential method deletes each potential factor variable from F to test whether the deleted variable is a true factor. If the MBIC estimate becomes $p - 1$, then it is put into the true factor set. However, it cannot be placed in the true factor set if the MBIC estimate remains as p. The number of identified factors in the first step (say, g) may be equal to or greater or smaller than p. All g variables are considered true factors when $g = p$. When $g < p$,

we need to identify $(p - g)$ more significant factors. The second step adds each potential factor from the $(q - g)$ variables in the matrix F to the identified factor set in the first step. If the MBIC estimator becomes $g + 1$, the added variable is considered a true factor. We repeat the same procedures until all p true factors are identified. When $g > p$, we need to delete $(g - p)$ variables from the g variables set. Thus, we eliminate the $(q - g)$ variables not in the true factor set from the matrix F and repeat the first step with the new matrix F until all of the p true factors are identified.

This sequential method may fail to end up with exactly p identified common factors and may produce more than p identified factors. The MBIC estimate is not necessarily equal to the number of aggregate variables with significant explanatory power but equals the number of *linear combinations of aggregate variables* that have explanatory power. For example, suppose $p = 4$ and three separate aggregate variables are true factors and a linear combination of two other variables in the matrix F are also true factors. The correct inference is that at least p factors (out of the q potential factors) are significant. Thus, the MBIC estimates the number of factors as four, but the sequential method may identify five relevant aggregate variables as common factors in this example.

For the specification of the attendance regression equation, we followed Ahn and Lee (2014) except for a few additions. We set the logarithm of annual average attendance per game (*Attpg*) as the dependent variable, and the regressor vector x_{it} includes team-specific variables, league-specific variables, and event variables. We denote this as regressor Set A.

$$x = \{Park\_i, Park\_s, Win\%, SLG, SOPG, L\_GU, L\_PU, L\_CSU, Dwar, Dstrike, DFA, D\text{exp}an, DOlymp, DCup\},$$

$$F = \{Trend, LPGDP, GDPG, ML\_RPG, ML\_HR, ML\_BA, ML\_SLG, ML\_SBPG, ML\_BBPG, ML\_SOPG\}.$$

Definitions of the variables are provided in Table 1. Team-specific variables include stadium effects (*Park_i*, *Park_s*), winning performance (*Win%*), and offensive (*SLG*) and pitching (*SOPG*) performance as well as league-average uncertainty of outcome measures (*L_GU*, *L_PU*, *L_CSU*). The uncertainty of game outcomes fits well conceptually with individual games, and the absolute deviation of the winning percentage from 0.5 can be a legitimate measure that controls for game uncertainty. However, the sample data in this study are annual, and thus the game uncertainty measure should represent the average game uncertainty of all daily games in a season (i.e., the league-wide average game uncertainty in a season). The conventional standard deviation is a well-developed statistic for *L_GU*. The measures of *L_PU* and *L_CSU* are the average difference in win percentages between the division winner and runners-up and the correlation between a team's winning percentage in the current season and the same team's average winning percentage over the previous three seasons, respectively. Smaller values of *L_GU*, *L_PU*, and *L_CSU* indicate greater uncertainty. The event variables are *Dstrick*, *DFA*, *Dexpan*, *Dwar*,

Table 1. Definition of Variables.

Variable	Definition
Team-specific variables	
Attpg	Average attendance per game
Park_s	Stadium capacity (number of seats)
Park_i	New stadium (four in the first season; three in the second season; two in the third season; one in the fourth season; zero after the fourth season).
Win%	Winning percentage
SLG	Slugging percentage
SOPG	Strikeouts per game that pitchers earned
L_GU	Game uncertainty of AL or NL[a]
L_PU	Playoff uncertainty of AL or NL[b]
L_CSU	Consecutive season uncertainty of AL or NL[c]
$\lvert Win\% - 0.5\rvert$	The absolute deviation of winning percentage from 0.5
PU(GB)	Games back to each division leader or a wild card leader
Dwar	Dummy variable for the years of World War I and World War II
Dstrike	Dummy variable for the years when a strike occurs
DFA	Dummy variable for the years when free agents are allowed
Dexpan	Dummy variable for the years when new teams are added
DOlymp	Dummy variable for the years of the Summer Olympic Games
DCup	Dummy variable for the years of the Federation Internationale de Football Association World Cup
Aggregate variables	
Trend	Time trend
PGDP	Per capita GDP
GDPG	GDP growth rate
ML_RPG	Average runs per game
ML_HR	Average home runs per game
ML_BA	Average batting percentage per game
ML_SLG	Average slugging percentage per game
ML_SBPG	Average stolen bases per game
ML_BBPG	Average bases on balls per game obtained
ML_SOPG	Average strikeout per game obtained
ML_GU	Average of L_GU
ML_PU	Average of L_PU
ML_CSU	Average of L_CSU

Note. AL = American League. NL = National League. GDP = gross domestic product.
[a]The conventional relative standard deviation of winning percentage. See Fort and Quirk (1995). [b]The average difference of win percentages between the division winner and runners-up. [c]The correlation between a team's winning percentage in the current season and the same team's average winning percentage over the previous three seasons.

Dolymp, and Dcup. The first three are relevant to MLB-specific information of the years when a strike occurs, when free agents are allowed, and when new teams are added, respectively. The next three variables nest the information for the years of World War I and World War II, the Summer Olympics, and the FIFA World Cup,

respectively. Aggregate variables as potential common factors are a time trend, macroeconomic variables (*LPGDP, GDPG*), and aggregate MLB statistics (*ML_RPG*, *ML_HR*, *ML_BA*, *ML_SLG*, *ML_SBPG*, *ML_BBPG*, *ML_SOPG*).

We also specify a different regressor vector x_{it}. We denote this as the regressor Set B. Instead of league-average uncertainty of outcome measures, this new regressor vector includes measures of outcome uncertainty that each home team faces. The absolute deviation of winning percentage from 0.5 controls for game uncertainty and has been used previously.[1] The games back, denoted by $PU(GB)$, is used as a team-specific measure of playoff uncertainty. A large number of $PU(GB)$ for a team implies that the chances that this team advances to the postseason are slim during the regular season, and therefore, the larger the $PU(GB)$, the less the playoff uncertainty. An advantage of adding team-specific measures of outcome uncertainty is that we are able to examine not only whether fan demand is influenced by home team-specific outcome uncertainty but also whether it is influenced by league-wide uncertainties. That is, league-wide uncertainty measures can be included in the potential common factor matrix F and we can examine whether they are identified as true factors.

$$x = \{Park\_i, Park\_s, Win\%, SLG, SOPG, |Win\% - 0.5|, PU(GB), Dwar, Dstrike, DFA, D\text{expan}, DOlymp, DCup\}.$$

$$F = \{Trend, LPGDP, GDPG, ML\_RPG, ML\_HR, ML\_BA, ML\_SLG, ML\_SBPG, ML\_BBPG, ML\_SOPG, ML\_GU, ML\_PU, ML\_CSU\}.$$

Sample Data and Empirical Results

The sample panel data set includes 34 MLB teams for the period 1904-2012. We treated teams with the same name, but with different home cities, as two different teams, and then it was a significantly unbalanced panel data set. We divided the sample data into two subsets. The first set covers the period from 1904 to 1957 when the Dodgers and Giants were located in New York. The second set covers the period from 1958 to 2012 after the Dodgers and Giants relocated to Los Angeles and San Francisco, respectively. This division and the following regression allowed us to analyze changes in fan preferences. The descriptive statistics of team-specific variables as well as aggregate variables are summarized in Table 2. Focusing on aggregate variables, there were several dramatic changes between the two periods. All three types of uncertainty measures displayed significant jumps in uncertainty. Game uncertainty, playoff uncertainty, and consecutive-season uncertainty increased, on average, by 28%, 17%, and 29% from the 1904-1957 to the 1958-2012 period, respectively. Home runs per game and strikeouts per game also increased significantly, by 107% and 68%, respectively.

Table 3 contains the estimation results of the attendance equation and the number of factors. Models 1 and 2 are the attendance regression with the regressor Sets A and B, respectively. Table 3 is similar to the main results of Ahn and Lee (2014). The estimated

Table 2. Descriptive Statistics for Team-Specific and Aggregate Variables.

	Mean			*SE*		
Variable	1904-1957	1958-2012	1904-2012	1904-1957	1958-2012	1904-2012
Team-specific variables						
Attpg	8,089	23,586	17,521	4,920	10,090	11,334
Park_s	32,618	47,886	41,926	15,657	10,303	14,675
Park_i	0.141	0.278	0.224	0.612	0.866	0.779
Win%	50.102	50.257	50.171	9.785	6.990	8.197
GB	21.205	13.078	16.301	15.912	11.326	13.902
SLG	37.020	39.707	38.649	4.159	3.336	3.904
SOPG	3.501	5.887	4.955	0.702	0.907	1.429
Aggregate variables						
PGDP	8,769	30,722	22,139	3,315	8,689	12,832
GDPG	3.413	2.997	3.157	6.950	2.174	4.659
ML_GU	2.535	1.819	2.098	0.354	0.274	0.466
ML_PU	0.047	0.039	0.042	0.024	0.015	0.020
ML_CSU	0.622	0.443	0.514	0.193	0.213	0.223
ML_RPG	4.388	4.410	4.401	0.515	0.363	0.428
ML_HR	0.432	0.893	0.714	0.233	0.155	0.294
ML_BA	26.567	25.939	26.182	1.465	0.750	1.127
ML_SLG	37.014	39.655	38.627	3.266	2.324	3.014
ML_SBPG	0.641	0.634	0.636	0.365	0.134	0.250
ML_BBPG	3.171	3.285	3.241	0.392	0.177	0.286
ML_SOPG	3.498	5.877	4.948	0.481	0.691	1.313

number of common factors was four in the first period in both Models 1 and 2. Models 1 and 2 include 10 and 13 aggregate variables in the potential factor matrix F, respectively, but the MBIC rank estimator provides the same number of factors from the two different matrices F. In the second period, the estimated ranks were 6 and 7 in Models 1 and 2, respectively. Thus, more common factors became significant in the modern era.

Table 4 reveals the identified common factors in the first period obtained by the sequential method. A time trend and per capita GDP were significant factors in attendance and among various performance statistics, only league-average stolen bases and strikeouts were identified as common factors. These results remain valid in Model 2 although it added three outcome uncertainty measures to the potential factor set. That is, competitive balance was not important in attending decisions. Table 3 implies that the MLB fans were influenced weakly by their home teams' playoff uncertainty according to Model 2 and Table 4 shows that MLB fans were not influenced by league-wide output uncertainties either. In summary, the fans in the early period were influenced mostly by their home teams' winning performance, and playoff uncertainty and aggregate variables of per capita GDP, running ability, pitching power, and something represented by a time trend. The factor loading estimates of stolen bases and strikeouts had negative signs, on average, while there

Table 3. Estimation of Attendance Equation.

	1904-1957				1958-2012			
	Model 1		Model 2		Model 1		Model 2	
Variable	Estimates	*t*-Values	Estimates	*t*-Values	Estimates	*t*-Values	Estimates	*t*-Values
Park_i	0.005	0.319	0.012	0.782	0.070	9.434	0.072	10.079
Park_s	−0.003	−1.886	−0.004	−2.09	0.005	3.349	0.005	3.524
Win%	2.364	16.408	1.383	3.318	1.988	16.677	1.773	8.577
SLG	0.009	1.259	0.011	1.530	0.013	3.266	0.011	3.040
AL × SLG	0.007	0.843	0.004	0.420	−0.007	−1.792	−0.005	−1.493
SOPG	−0.013	−0.353	−0.033	−0.852	0.005	0.336	0.017	1.094
AL × SOPG	0.008	0.169	0.028	0.612	0.008	0.412	−0.010	−0.489
L_GU	0.011	0.603			−0.049	−2.875		
L_PU	−1.876	−7.994			−0.179	−0.742		
L_CSU	0.053	1.321			−0.048	−2.203		
|Win% − 0.5|			0.215	1.310			−0.311	−2.173
PU(GB)			−0.007	−2.527			−0.001	−1.210
Dwar	−0.227	−7.467	−0.202	−6.754				
Dstrike					−0.338	−7.678	−0.326	−7.551
DFA					0.162	4.275	0.125	3.728
Dexpan					−0.005	−0.25	0.032	1.510
DOlymp	−0.029	−1.645	−0.048	−2.633	−0.005	−0.416	−0.002	−0.156
DCup	−0.166	−6.250	−0.196	−7.527	−0.006	−0.529	−0.011	−1.079
MBIC rank	4		4		6		7	
Adj. R^2	.871		.873		.846		.859	

Note. The *t*-values are computed with standard errors robust to conditional heteroskedasticity.

was variation in the estimates across teams. That is, fans in the first period did not prefer games with many stolen bases and strikeouts, on average.

Table 5 reveals the identified common factors in the second period. Neither Model 1 nor Model 2 produced a set of identified common factors, but two or three sets, respectively. As discussed above, our sequential method may identify aggregate variables as common factors more than p. This occurs if several aggregate variables form a common factor as a linear combination. In Model 1, there are six common factors and the five aggregate variables were identified as separate common factors and two variables form a common factor. Thus, the sequential method produced two sets of aggregate variables as identified common factors. The five factors among the six identified common factors are relevant to offensive performances and the last factor is likely the combination of time trend and per capita GDP. In Model 2, league average outcome uncertainties are included as potential factors and, in fact, game uncertainty and consecutive season uncertainty were identified as common factors. As shown in Model 1, the combination of time trend and per capita GDP is likely to be a common factor influencing attendance at MLB

Table 4. Identification of Common Factors Using Sequential Method: 1904-1957.

Model 1		Model 2	
Variable	Common Factors	Variable	Common Factors
Trend	Trend	Trend	Trend
PGDP	PGDP	PGDP	PGDP
GDPG	ML_SBPG	GDPG	ML_SBPG
ML_RPG	ML_SOPG	ML_GU	ML_SOPG
ML_HR		ML_PU	
ML_BA		ML_CSU	
ML_SLG		ML_RPG	
ML_SBPG		ML_HR	
ML_BBPG		ML_BA	
ML_SOPG		ML_SLG	
		ML_SBPG	
		ML_BBPG	
		ML_SOPG	

Table 5. Identification of Common Factors Using Sequential Method: 1958-2012.

Model 1			Model 2			
Variable	Common Factors		Variable	Common Factors		
Trend	Trend	PGDP	Trend	Trend	Trend	PGDP
PGDP	ML_RPG	ML_RPG	PGDP	PGDP	ML_GU	ML_GU
GDPG	ML_HR	ML_HR	GDPG	ML_GU	ML_CSU	ML_CSU
ML_RPG	ML_BA	ML_BA	RSD	ML_CSU	ML_RPG	ML_RPG
ML_HR	ML_SBPG	ML_SBPG	PU	ML_BA	ML_BA	ML_BA
ML_BA	ML_BBPG	ML_BBPG	CSU	ML_SLG	ML_SLG	ML_SLG
ML_SLG			ML_RPG	ML_SOPG	ML_SOPG	ML_SOPG
ML_SBPG			ML_HR			
ML_BBPG			ML_BA			
ML_SOPG			ML_SLG			
			ML_SBPG			
			ML_BBPG			
			ML_SOPG			

and offensive variables of league-average runs and batting percentage were also significant. League-average slugging percentage and strikeouts, which were not significant in Model 1, were identified as common factors. Considering that both home runs and slugging percentage are relevant to hitters' slugging ability, league-average strikeouts was the only factor not found in Model 1.

For further discussion, we will focus on the results of Model 2 because there were more aggregate variables in a potential factor matrix and this larger factor matrix enables us to examine whether league-wide outcome uncertainties influenced individual teams' attendance. An intriguing fact when we calculated the factor loading estimates was that the average estimate of slugging was positive whereas that of batting average was negative. The estimate of strikeouts also had a negative sign, on average, in the modern era. These empirical results imply that fans in the second period preferred offensive performance, but in a particular way. They preferred offensive games with slugging performance instead of offensive games with many single hits, but no extra base hits. The causes of the fan preference movement toward slugging would be an intriguing topic for further study. One suggestion is the home run race episode in the 1961 MLB season. The race between Maris and Mantle to break Ruth's home run record drew not only the interest of Yankees fans but also nationwide interest.

The factor loading estimates of game uncertainty and consecutive season uncertainty (*ML_GU*, *ML_CSU*) were negative, on average. This implies that the improvement of the uncertainties increased attendance, on average. Calculating the changes in attendance due to increase in uncertainties, we found that improvement in *ML_GU* (from 2.535 in the first period to 1.819 in the second period) and *ML_CSU* (from 0.622 to 0.443) helped MLB to increase attendance per game, by 7.9% and 2.3%, respectively, $.079 = (1.819 - 2.535) \times (-0.111)$ and $0.023 = (0.443 - 0.622) \times (0.127)$, where -0.111 and -0.127 are the averages of the factor loading estimates, respectively. These two facts together suggest that fans began to prefer uncertainties in the modern era and that the uncertainties increased in the modern era, helping to draw more attendance per game, on average, in the second period by 10.2% than in the first period.

Comparing the results of the two periods, there are several intriguing findings. First, MLB fans have been influenced consistently by income and a factor represented by a time trend. Second, league-wide outcome uncertainties became significant in the modern era. None of the three types of uncertainty measures were identified as a significant factor, and only team-specific playoff uncertainty was statistically significant in attendance in the first period. However, both league-wide game uncertainty and consecutive season uncertainty (*ML_GU*, *ML_CSU*) were significant in addition to home teams' game uncertainty. This implies that MLB fans in the modern era consider not only their home teams' characteristics but also the away teams' characteristics. For example, considering the significance of *ML_CSU*, the decision to attend a baseball game is influenced by whether an incoming away team is a perennial winner.

Third, league-wide offensive performance became significant factors in the second period. This empirical result is additional evidence that the characteristics of away teams have become significant determinants of attendance during the modern era. In fact, previous empirical works have found the same results for different team sports: the English Premier League in Buraimo and Simmons (2008) and the

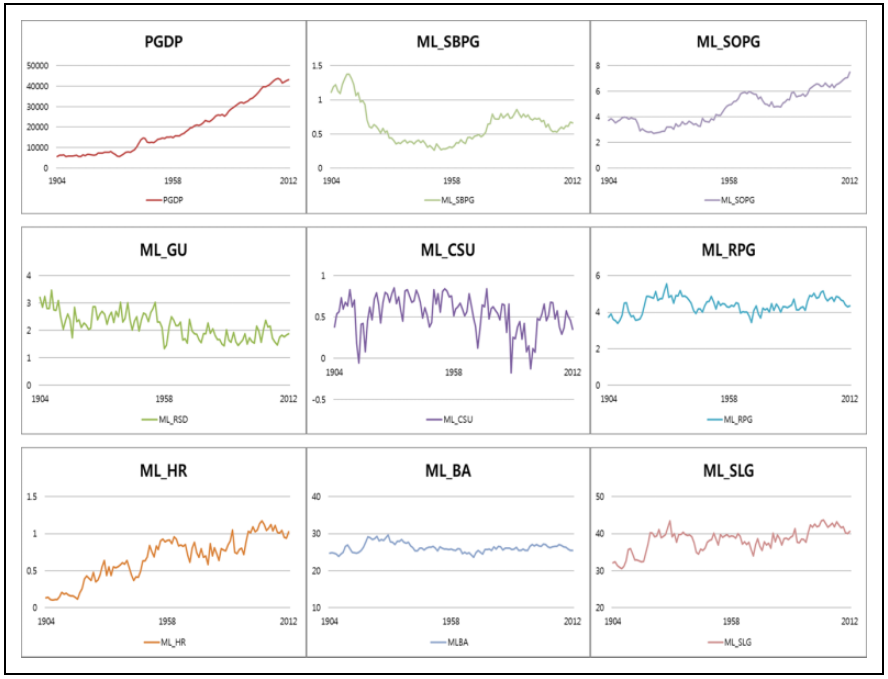

Figure 1. Temporal fluctuations of common factors.

National Hockey League in Coates and Humphreys (2012). They analyzed only five to six seasons during the 2000s. Our result shows that this finding was also true during earlier periods, although not in 1901-1957. Groothuis, Rotthoff, and Strazicich (2015) used time-series tests with structural changes in several MLB performance statistics (home runs, slugging percentage, batting average, and runs batted in) over 140 seasons. They found that structural breaks in slugging percentage in 1921 and 1992, and suggested the free swing (Babe Ruth) era and the modern steroid era as potential causes, respectively. Our estimation results, that league-wide slugging performance became a significant common factor in the modern era in addition to the fact that home teams' slugging ability became a determinant of attendance, may be relevant to the structural break in 1992 and the use of steroids. Players might notice a shift in fan preferences, so that slugging performance became a more valuable ability in the modern era, and then there was more incentive to improve player power.

Figure 1 shows nine aggregate variables that were identified as common factors in either the 1904-1957 or 1958-2012 period. The average number of stolen bases was a common factor in the first period but became insignificant in the second period; it was greater than once per game in 1900-1920 but decreased to and stayed at about 0.7 in 1970-2012. Strikeouts showed a consistently increasing trend

although it was a significant common factor, but a negative factor loading, on average, during the whole sample period. However, slugging percentage was significant only in the second period and the temporal pattern was constant and increasing in the first and second periods, respectively. That is, both strikeouts and slugging percentage increased consistently during the second period. The fact does not indicate that hitters' ability overpowered pitchers' ability or vice versa. Batting average was a significant common factor and its factor loading was negative, on average, in the second period; it decreased from 1930 to 1950, but since then, it stayed more-or-less constant while slugging percentage increased. It is intriguing that hitters have been successful in develop their skills at slugging, which is significantly positive in fan preferences. Other common factors that were significant only in the second period were offensive performance in terms of runs per game (*ML_RPG*) and two uncertainty measures (*ML_GU*, *ML_CSU*). The temporal changes in league-average game uncertainty and consecutive uncertainty also show that both uncertainties were greater in the second period although the decline in *ML_CSU* after 1958 was slim. Note that the lower the *ML_GU* or *ML_CSU*, the greater the uncertainty. Coincidentally, all of the four common factors that were insignificant in the first period and became significant in the second period, then improved in the second period.

Further Discussion and Concluding Remarks

This article used a panel data set with a long time series, more than 100 years, to identify common factors in MLB attendance. This was achieved by applying a panel data model with observed common factors and a sequential test method. A major reason why previous studies on attendance could not use a panel data sample with a long time series has been incomplete observations of ticket prices that would be expected to be an important regressor, although there are complete data on team attendance and other team characteristics, such as winning performance. Ahn and Lee (2014) assumed that a member team of MLB determines its ticket price based on information of common factors that other member teams also take account of when setting their ticket prices. Macroeconomic conditions, such as GDP, could be a candidate common factor. However, each team could also consider characteristics of its territory market and then there have been heterogeneous prices across member teams. Our econometric model includes an individual effects term and 13 observed potential common factors in the case of Model 2. The number of significant common factors was estimated to be seven in the period of 1958-2012; the identified factors were a time trend, per capita GDP, outcome uncertainties, several offensive statistics, and strikeouts. This econometric model can be legitimate enough to produce consistent estimates if ticket prices are a function of linear combinations of common factors. Considering wide variation in ticket prices across member teams, individual effects may be correlated with ticket prices, but they are not likely to satisfy the time-invariant restriction. Thus, the time trend and/or per capita GDP may be common factors relevant to ticket prices.

This study compared estimation results of two different subsets of the data from two time periods (1904-1957 and 1958-2012). Ahn and Lee (2014) found that MLB fan preferences were simple in the early years, when only winning mattered. However, fan preferences became influenced by various factors in later years. Team-specific variables of outcome uncertainty, stadium quality, and batting performance started to significantly influence attendance. This study also found that MLB fan preferences were simple in the early years with respect to common factors and then became multifaceted in later years, because the number of significant common factors increased from four to seven. A time trend and per capita GDP were significant in the whole sample period, but outcome uncertainties and offensive performance (in particular, slugging performance) became newly significant common factors influencing attendance in later years.

Coincidentally, average runs per game (*ML_RPG*), average batting percentage (*ML_BA*), and average slugging percentage (*ML_SLG*), which were not significant in the early years but became significant common factors in later years, displayed a temporal trend of decline (or at least were constant) when insignificant and increased when significant. This may be a reflection that players and teams understood changes in fan preferences and developed skills that fans preferred. The MLB may have successfully implemented rules or created an environment that promoted playing skills that fans prefer. The emergence of league-wide offensive performance as a significant common factor in the modern era has important policy implications. It implies fans consider not only home team characteristics but also the offensive characteristics of away teams. More elaborate league business rules from scheduling to strike rulings should be carefully examined to provide attractive away teams and to promote slugging performance. The importance of offensive performances in attendance drawing also calls for further investigation of the relationship between attendance and league-wide offensive performance. Previous studies, and this article, have focused on average offensive performance, but examinations of higher moments of offensive performances would provide additional valuable information. According to Groothuis, Rotthoff, and Strazicich (2015), there are clear temporal patterns in standard deviations of home runs and batting averages in MLB. The variation in home runs among players has increased consistently, while variation in batting average decreased over the period of 1871-2010. An analysis of the relationship between fan preferences and variation in offensive performance among players can be examined in future research.

Our empirical finding that fans took account of home teams' outcome uncertainty but also league-wide uncertainties in later years deserves more discussion. This is direct empirical evidence that the maintenance of a competitive balance within a reasonable range is a critical issue for the profitability of MLB in later years. Previous empirical results on the uncertainty of outcome hypotheses have been mixed (Coates et al., 2014). Our findings suggest that the hypothesis was rejected in early years but could not be rejected in later years. This may explain the mixed results of previous studies, which may have used samples with different time spans.

It also implies that fans are influenced by away teams' outcome uncertainty. This is consistent with previous findings that modern fans also consider offensive performances of away teams. Additional empirical evidence consistent with this finding is the variation in fan loyalty. According to the fan loyalty estimates of Ahn and Lee (2014), the difference between the maximum and minimum fan loyalties in the first period was 1.22, but it declined to 0.66 in the second period. The standard deviations of fan loyalties across member teams were 0.34 and 0.19 in the first and second periods, respectively. The fact that characteristics of away teams became significant determinants of attendance in the second period reduced the influence of local characteristics and then the disparity in fan loyalty lessened in later years. What caused changes in fan preferences such that fans considered not only home team characteristics but also characteristics of away teams is an important question. We suggest that the development of transportation, increased mobility in residences, and developments in media are potential causes. It is not uncommon to find fans rooting for away teams in MLB ballparks today. The development of transportation has reduced the time cost of following home teams when they go away. Some fans who moved from another city may greet their old teams when they are in town. Increased mobility in residence makes this a more common occurrence. Certainly, there should be more mobility in recent years than in the years before 1958. The development of media has provided fans with more information about other teams and fans have greater knowledge of incoming away teams, which may influence their attending decisions. The relationship between fan preferences and away team characteristics deserves further investigation in future studies.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the National Research Foundation of Korea Grant funded by the Korean Government (NRF-2016S1A2A2912).

Note

1. For outcome uncertainty of an individual match, Baimbridge, Cameron, and Dawson (1996) used the absolute difference in league rankings, whereas Wilson and Sim (1995) used the absolute difference in points.

References

Ahn, S. C., Horenstein, A. R., & Wang, N. (2013). Beta matrix and common factors in stock returns (Working Paper). Coral Gables, FL: University of Miami.

Ahn, S. C., & Lee, Y. H. (2014). Major League Baseball attendance: Long-term analysis using factor models. *Journal of Sports Economics*, *15*, 451–477.

Ahn, S. C., Lee, Y. H., & Schmidt, P. (2013). Panel data models with multiple time-varying individual effects. *Journal of Econometrics*, *174*, 1–14.

Bai, J. (2009). Panel data models with interactive fixed effects. *Econometrica*, *77*, 1229–1279.

Baimbridge, T. M., Cameron, S., & Dawson, P. (1996). Satellite television and the demand for football: A whole new ball game? *Scottish Journal of Political Economy*, *43*, 317–333.

Buraimo, B., & Simmons, R. (2008). Do sports fans really value uncertainty of outcome? Evidence from the English Premier League. *International Journal of Sport Finance*, *3*, 146–155.

Coates, D., & Humphreys, B. (2012). Game attendance and outcome uncertainty in the National Hockey League. *Journal of Sports Economics*, *13*, 364–377.

Coates, D., Humphreys, B., & Zhou, L. (2014). Reference-dependent preferences and loss aversion and live game attendance. *Economic Inquiry*, *52*, 956–973.

Fort, R., & Quirk, J. (1995). Cross-subsidization, incentives, and outcomes in professional team sports leagues. *Journal of Economic Literature*, *23*, 1265–1299.

Groothuis, P. A., Rotthoff, K. W., & Strazicich, M. C. (2015). Structural breaks in the game: The case Major League Baseball. *Journal of Sports Economics*. doi:10.1177/1527002515593113

Keloharju, M., Linnainmaa, J. T., & Nyberg, P. (2016). Return seasonalities. *Journal of Finance*, *71*, 1557–1590. doi:10.1111/jofi.12398.

Wilson, P., & Sim, B. (1995). The demand for Semi-Pro League Football in Malaysia 1989–1991: A panel data approach. *Applied Economics*, *27*, 131–138.

Author Biography

Young H. Lee, PhD, is a professor in the Department of Economics at Sogang University. His research fields include econometrics (panel data models and productivity analysis) and sports economics.