Asia Pacific Journal of Information Systems
Vol. 25 No. 3 (SEPTEMBER 2015), 563-578

ISSN 2288-5404 (Print) / ISSN 2288-6818 (Online)
http://dx.doi.org/10.14329/apjis.2015.25.3.563

Using the Hierarchical Linear Model to Forecast Movie Box-Office Performance: The Effect of Online Word of Mouth

Jongmin Park[a], Yeojin Chung[b,*], Yoonho Cho[c]

[a] *Graduate Student, Graduate School of Business Administration, Kookmin University, Korea*
[b] *Assistant Professor, College of Business Administration, Kookmin University, Korea*
[c] *Professor, College of Business Administration, Kookmin University, Korea*

ABSTRACT

Forecasting daily box-office performance is critical for planning the distribution of marketing resources, and by extension, maximizing profits. For certain movies, the number of viewers increases rapidly at the beginning of their theatrical run, and the increments slow down later. Other movies are not popular in the beginning, but the audience sizes grow rapidly afterward. Thus, the audience attendance of movies grow in different trajectories, which are influenced by various factors including marketing budget, distributors, directors, actors, and word of mouth. In this paper, we propose a method for predicting the daily performance trajectory of running movies based on the hierarchical linear model. More specifically, we focus on the effect of online word of mouth on the shape of the growth curves. We fitted the mean trajectory of the cumulative audience size as a cubic function of time, and allowed the intercept and slope to vary movie-to-movie. Moreover, we fitted the linear slope with a function of online word of mouth predictors to help determine the shape of the trajectories. Finally, we provide performance predictions for individual movies.

Keywords: Box-Office Performance, Hierarchical Linear Model, Growth Curve Model

Ⅰ. Introduction

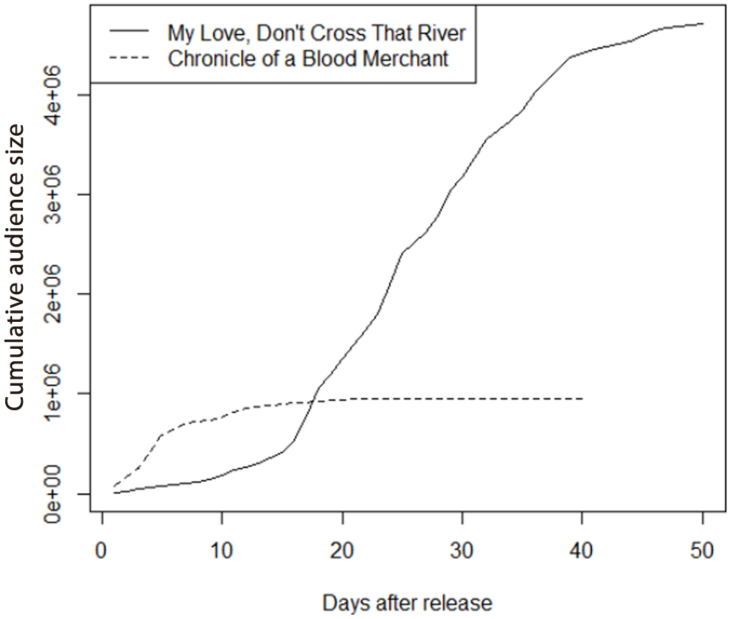
In the winter of 2014 - 2015, a famous Korean actor and film director, Jung-woo Ha, released a movie entitled "Chronicle of a Blood Merchant." Before its release, this film had drawn significant attention from the public, as it was the first movie Ha directed, featured a star-laden cast, and depicted the bestselling Chinese novel with the same title. On its opening weekend, more than 400,000 people viewed the film. Even with the strong first weekend, the number of people that went to see the movie grew slowly after its opening week (see <Figure 1>). In contrast, only about 50,000 moviegoers saw the documentary film "My Love, Do not Cross That River" on its opening weekend. Its audiences grew slowly at first but rapidly between 20 and 40 days after the film's release. Ultimately, nearly five million

*Corresponding Author. E-mail: ychung@kookmin.ac.kr Tel: 8229105614 moviegoers saw "My Love, Do not Cross That River," far outpacing the less than one million that saw "Chronicle of a Blood Merchant." Given the marked lack of star-power in the film's cast, the success of "My Love, Do not Cross That River" is largely attributable to word of mouth (WOM). While it had an online review score of 8.9 out of 10 on *Daum Movie*, "Chronicle of a Blood Merchant" had only a 6.7. As evidenced by these examples, audience attendances adhere to different trajectories that are influenced by various factors, including WOM, marketing budget, distributors, directors, and actors. Recognizing these trajectories is critical to forecast a film's box-office performance.

Because of its unpredictability, many scholars consider forecasting box-office performance to be a challenging endeavor. Still, many researchers have attempted to model movie success using statistical methods because forecasting daily box-office performance is critical for planning the distribution of marketing resources, and by extension, maximizing profits. Many researchers have employed multiple regression models (Asur and Huberman, 2010; Basuroy et al., 2003; Elisashberg and Shugan, 1997; Liu, 2006; Park and Song, 2012; Ravid, 1999) and Kim and Hong(2013) used lasso and ridge regression to avoid multicollinearity among predictors.

Some authors have transformed the prediction problem into a classification problem by discretizing movie success. In doing so, they have typically sought to predict a film's success before it is released in theaters. These researchers have built classification models including neural networks (Sharda and Delen, 2006), multi-layer back-propagation neural networks (Zhang et al., 2009), multinomial logit regression models (Kim and Hong, 2011), and Bayesian selection models (Lee and Chang, 2006).

Although the aforementioned studies have been principally geared towards identifying factors that contribute to a film's financial success at a given point in



<Figure 1> Cumulative Number of Audiences for Two Movies Released Recently

time (mostly its first weekend or at the end of its theatrical run), some studies have explored the dynamic structure of movie success by repeatedly fitting a regression model for every weekend a movie is in theaters (Liu, 2006; Park and Song, 2012). They have explored changes in a movie's box office success through graphical or numerical summaries, but have not considered correlations of a movie's performances at different time points in time. Some authors developed and employed models based on diffusion theory with an exponential distribution (Jedidi et al., 1998; Swami et al., 1999), a gamma distribution (Ainslie et al., 2005; Sawhney and Eliashberg, 1996), and the Bass diffusion model (Dellarocas et al., 2007). But these models have restriction to fit the trajectory of the growth in a given parametric form. Autoregressive models (Rui et al., 2013) and simultaneous equation models (Duan et al., 2008) are alternatives to capture the dynamic nature of box-office performance, but they are not appropriate to describe the growth patterns.

In remedy of the shortcomings of past work in this domain, we propose a method of capturing the growth pattern of the cumulative audience attendance using a hierarchical linear model and predict a running movie's daily performance based on the estimated growth curve. Hierarchical linear models, also known as "linear mixed models" or "multilevel linear models" are widely used for longitudinal data, cross-sectional data on subjects nested in groups. Box-office data is a typical example of longitudinal data since daily box-office returns are nested within a given movie. We use a time variable (days after release) and its polynomial functions as explanatory variables to capture the outcome variable's (cumulative size of the audience) pattern of growth in the day-level model. We allow the slope of the time variable and the intercept to vary among movies on the basis of randomness and various predictors (e.g., directors, distributors, genre, and WOM) observed at the movie-level. We also include daily characteristics predictors (e.g., day of the week) in the day-level model to control for the weekend effects on the box office.

Through the use of these methods, this study offers several important contributions on the literature related to box office prediction. First, our hierarchical linear model (also called a growth curve model) explicitly and systematically models the shapes of the cumulative audience sizes' trajectories over time. Although diffusion models are also suitable for modeling the growth of total moviegoers, they restrict the trajectory of the growth curve in a specific parametric form (e.g., exponential, gamma distribution curves, or Bass model.) This reduces the model's flexibility to explain the various types of the growth curve compared to the polynomial function of the time variable.

Second, the proposed model enables us to observe the effect of movie-level covariates on the shape of the growth curve. We are particularly interested in how online WOM influences the trajectory of box-office performance. Several researchers regard WOM as one of the most influential factors of movie sales (Chintagunta et al., 2010; Dellarocas et al., 2007; Duan et al., 2008; Lee et al., 2013; Liu, 2006; Rui et al., 2013). In this paper, online WOM variables directly define the shape of the growth trajectory curve related to box-office performance. By expressing the slope and the intercept of the time variable as functions of online WOM variables, we can explore the impact of these variables on the mean level and speed of the growth of the cumulative number of audience members.

Third, this model considers the heterogeneity of the growth curves across movies by including random effects to the intercept and slopes associated with the time variables. As a result, we account for uncertainty in predicting the success of new movies. Without incorporating random effects, standard errors intrinsic to the model are underestimated, thereby causing the researcher to underestimate his/her error in making predictions (Draper, 1995; Gelman and Meng, 1996).

To address the issues described above, we have organized the paper into a series of interrelated and sequential sections. In Section 2, we define the nature of the problem and the variables we use in the prediction model. Following, in Section 3, we briefly explain the growth curve model as a special case of the hierarchical linear model. We then present the results of the fitted model and prediction in Section 4, and offer some concluding remarks in Section 5.

Ⅱ. Data Preparation

2.1. Data Collection

In this paper, we use the box office data available at the time of prediction to forecast the daily performance of running movies in future. We assume a hypothetical case in which a target movie has been released a week before, and so box office data are available for the first week of release (Day 0 to Day 6.) Using these data, we propose a model for predicting a movie's daily performance for the second week (Day 7 to Day 13.) Naturally, our model can be generalized to predict performance of Day T using the box office data for Day 0 to Day t^* (< T.) In this paper, we focus on predicting performance when T = 7, 8, ⋯, 13 and t^* = 6.

We collected box office data for 777 movies released between January 2009 and September 2014 from the Korea Film Council.[1] To measure online WOM, we collected review scores and the volume of public comments from the *Daum Movie Review*[2] by web crawling method. Because noncommercial and minor movies are beyond the analytical scope of this study, we excluded films with cumulative audiences of fewer than 100,000 people 30 days after release.

We use cumulative audience size as a proxy measure of box-office performance because of its close association with revenue (Kim et al., 2010; Kim and Hong, 2013; Scott, 1994) and the ease with which the data could be collected relative to other types of financial information. Rather than predicting the sizes of daily audiences, we seek to forecast the cumulative sizes of audiences, as doing so allows us to explore differences in the growth curves of the cumulative audiences and to identify factors that affect growth trajectories.

2.2. Predictor Variables

To predict a film's financial success, we consider a number of movie-level predictors including movie characteristics, external factors, and online word of mouth. In the day level, the number of days after release is used as the time variable to define the shape of the growth curves and the day of the week is also included in the model. The variable definitions appear in <Table 1>.

2.2.1. Days after Release

Days after release is the time variable (t) that frames the trajectory of the response variable. Because we are primarily interested in the average patterns of response variable's growth, we include a polynomial function of t in the hierarchical model (see Section

[1] http://www.kobis.or.kr/kobis/business/stat/boxs/findDailyBoxOfficeList.do.

[2] http://movie.daum.net/review/netizen_point/movieNetizenPoint.do.

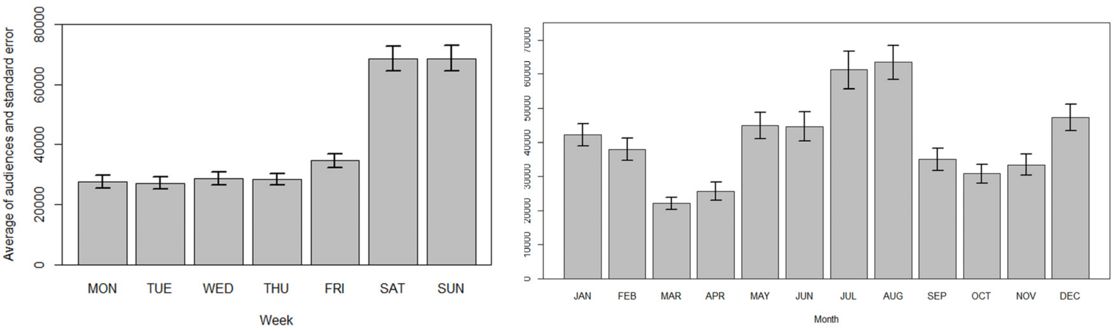
3.) To avoid multicollinearity among t and its polynomial functions, we have mean-centered the *days after release* variable. That is, it originally ranged from 0 days to 13 days after release, but we have transformed it to range from -6 to 6.

2.2.2. Day of the Week

Because more people tend to go to the movies on weekends relative to weekdays, the *day of the week* is a key factor in forecasting daily audience

<Table 1> Predictor Variable Descriptions

	Variable		Variable Description
Day level (level 1)	Days after release (t)		The number of days since the film was released
	Day of the week		Categorizes the day as falling on a weekend or weekday: (1) Sat and Sun, (2) Mon to Fri
Movie level (level 2)	Movie characteristics	Ave. num. of aud. of directors	Average total audience size for director's past films
		Genre	Represents the film's genre: (1) animation, (2) documentary, (3) thriller, (4) melodrama/romance, (5) horror, (6) drama, (7) action/adventure, (8) comedy, (9) historical drama
		Rating	Age restrictions imposed by the Korean Media Rating Board: suitable for (1) all ages, (2) 12 years and over, (3) 15 years and over, (4) 18 years and over
		Nationality	Nationality of the film: (1) Korea, (2) USA, (3) others
	External factors	Current cum. num. of aud.	Cumulative audience size at t = 6
		Ave. num. of aud. of distributors	Average cumulative audience size of the films that each company distributed
		Month	Categorizes the month the film is released into one of three time frames: (1) Jul and Aug, (2) Jan, Feb, May, Jun, and Dec, (3) Mar, Apr, Sep, Oct, and Nov
	Online word of mouth	Num. of comm. before release (Count.before)	Number of online comments during the week before the film's release
		Num. of comm. after release (Count.after)	Number of online comments during the week after the film's release
		Review score before release (Score.before)	Average online review score during the week before the film's release
		Review score after release (Score.after)	Average online review score during the week after the film's release



<Figure 2> Average Audience Size by Day of the Week and Month

sizes. The left panel of <Figure 2> summarizes the average audience size for each day of the week. Because these averages do not differ critically between Saturday and Sunday, or Monday to Friday, we used two classifications to define the days on which an individual attends a movie: Saturday and Sunday (1) and Monday through Friday (2).

2.2.3. Average Audience Size by Distributors/Directors

Film distributors and directors are important factors for predicting a film's success. However, the numbers of distributors (81) and directors (781) in the sample are too large to include these factors as dummy variables in the model. Instead, we calculated average of total audience sizes for the films that each company distributed or each director created, and used this as a numerical predictor. For movies with multiple directors, we used the first listed director as that film's representative.

2.2.4. Genre

Many researchers have shown that a film's genre is closely related with its box-office performance. In the United States, films in the comedy, science fiction, and horror genres have tended to perform well financially (Litman and Kohl, 1989; Sochay, 1994; Wyatt, 1991). In contrast, drama films tend to be negatively related to box-office success (Chang and Ki, 2005; Litman and Kohl, 1989). To account for variation in a movie's performance as a function of its genre, we classified movies into nine groups (Jang et al., 2009; Park et al., 2011): animation (14.7%), documentary (1.0%), thriller (4.5%), melodrama/romance (6.5%), horror (6.7%), drama (22.7%), action/adventure (34.4%), comedy (8.2%), and historical drama (1.3%), and included each film's genre category into the model.

2.2.5. Rating

A movie's rating indicates its age restrictions as determined by the Korean Media Rating Board. Since it specifies the potential market size of a movie, a rating has an important influence on movie's financial success. For the purposes of this study, films were categorized and coded as "suitable for all audiences" (1), "suitable for a person aged 12 years and over" (2), "suitable for a person aged 15 years and over" (3), and "suitable for a person aged 18 years and over" (4).

2.2.6 Nationality

Nationality of a movie indicates the nation in which the film's production studio is located. In our sample, 36.7% of movies are Korean, and 47.6% are American. Given this disparity, we coded each movie's nationality to account for variation as a function of it. Specifically, we coded Korean movies as 1, American movies as 2, and other movies as 3.

2.2.7. Current Cumulative Audience Size

The movies' cumulative audience sizes at the time of prediction represent their current success, which affects future success. Because we seek to predict a film's second-week performance on the last day of the first week, we use the cumulative audience size at $t = 6$ as a predictor.

2.2.8. Month

Average audience sizes vary across months (see the right panel of <Figure 2>). For example, because college students represent a major customer group in Korea's movie market, the summer and winter break seasons tend to have larger audiences than the spring and fall semesters. Therefore, we grouped the movies which were released in months with similar average audiences. These groupings are July and August (1), January, February, May, June, and December (2), and March, April, September, October, and November (3).

2.2.9. Number of Online Comments Before/After Release

The number of online comments about a movie reflects public interest in it. Regardless of the comments' inherent sentiment (i.e., positive vs. negative), interest alone can influence a movie's success (Kim and Lee, 2009). As such, we counted the number of comments about a movie that appeared online in the week prior to its release and in the first week after its release (Days 0 to 6). Since the distribution of the number of online comments is skewed to the right, we use the log transformation of the number of comments plus one (to avoid the negative infinity) as a predictor.

2.2.10. Online Review Score Before/After Release

Online audience review scores gauge the effect of word of mouth on the movie's performance (Liu, 2006). Positive word of mouth can increase audience size rapidly, even for movies to which little attention was paid before release. To account for this variable, we averaged the online review scores on the *Daum Movie Review* website for a week prior to a film's release and a week after its release.

<Table 2> Summary Statistics and Correlation Matrix of Key Variables

	n	Mean	sd	Median	Min	Max
Cumulative number of audience	10878	561002.61	803318.57	267753	254	11774754
Ave. num. of aud. of directors	777	1191964.44	1490776.14	638704	30684	9748737
Ave. num. of aud. of distributors	777	1194208.04	601895.62	1124020.61	79149.5	3883516
Score.before	777	8.23	1.63	8.62	0	10
Count.before	777	47.56	84.54	24	0	1608
Score.after	777	7.59	1.45	7.83	0	10
Count.after	777	306.81	652.57	154	0	14191

	Ave. aud. size of directors	Ave. aud. size of distributors	Score.before	Count.before	Score.after	Count.after
Ave. aud. size of directors	1.000	0.368	0.064	0.327	0.084	0.529
Ave. aud. size of distributors	0.368	1.000	0.018	0.138	0.018	0.214
Score.before	0.064	0.018	1.000	0.024	0.454	0.010
Count.before	0.327	0.138	0.024	1.000	0.101	0.755
Score.after	0.084	0.018	0.454	0.101	1.000	0.037
Count.after	0.529	0.214	0.01	0.755	0.037	1.000

Ⅲ. Hierarchical Linear Models

For 100 randomly selected movies, cumulative audience sizes are plotted over days after release in <Figure 3>. While the individual trajectory lines are gathered near the horizontal axis in its original scale (left), the log-transformed lines (right) are distributed more evenly and have a nonlinearly increasing pattern. To ease interpretation of our data and capture the nonlinear nature of the response variable in the day-level model, we use log-transformed cumulative audience size (right panel of <Figure 4>) as the response variable and include the polynomials of time t in the model.

Specifically, the day-level model is a function of the day-level independent variables outlined in <Table 1>. This model is expressed as

$$y_{ij} = \beta_{0j} + \beta_{1j} t_{ij} + \beta_2 t_{ij}^2 + \beta_3 t_{ij}^3 + \beta_3 d_{ij} + \epsilon_{ij},$$

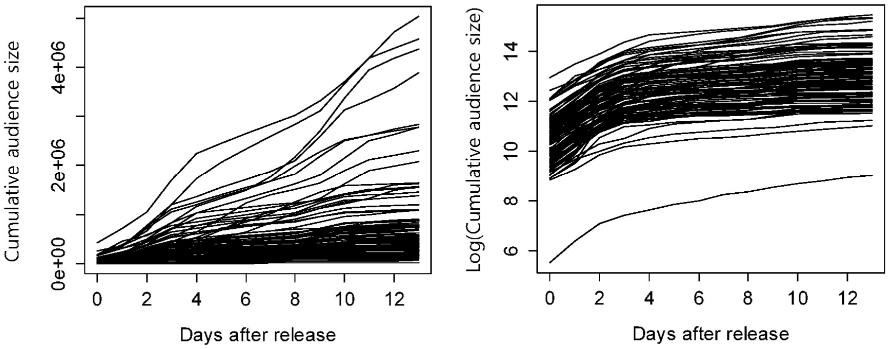
$$i = 0,\dots,13,\ j = 1,\dots,777$$

$$\epsilon_{ij} \sim N(0, \sigma_\epsilon^2)$$

where y_{ij} is the cumulative audience size on the i-th day after the release of the j-th movie, $t_{ij} = i - 6$ is the mean-centered number of days after the release of the j-th movie, d_{ij} is the day of the week, and the β's are regression coefficients. The random error ϵ_{ij} is the residual or error term for the i-th day of the j-th movie, assumed to be independent of the residuals for both a different day of the same movie and a day of a different movie. In the day-level model, the growth of cumulative audience size is expressed as a cubic function of t_{ij}. Though terms of a higher order than t_{ij}^3 can also be included model, they did not substantially improve the residual plot (not shown in this paper.) Because the residual plot in <Figure 4> does not have clear pattern, we decided that the day-level model should be a cubic function of t_{ij} for the sake of simplicity and parameter parsimony. We allowed the intercept β_{0j} and the coefficient of the linear term β_{1j} to vary among the movies to describe different growth patterns attributable to movie-level characteristics (particularly WOM).

The movie-level model explains differences in variation of β_{0j} and β_{1j} among the movies with the movie-level covariates and random error, given by

$$\beta_{0j} = \gamma_{00} + \gamma_{01} x_{1j} + \cdots + \gamma_{0,11} x_{11,j} + \eta_{0j}$$
$$\beta_{1j} = \gamma_{10} + \gamma_{11} x_{1j} + \cdots + \gamma_{14} x_{4j} + \eta_{1j}$$



<Figure 3> Individual Trajectory of Cumulative Audience Sizes and Its Log Transformation

where $x_{1j}, \cdots, x_{4j}$ are the online WOM variables; $x_{5j}, \cdots, x_{11,j}$ are the other movie-level variables from <Table 1> (excluding WOM variables); γ_{00}, $\cdots$, $\gamma_{0,11}, \gamma_{10}, \cdots, \gamma_{14}$ are regression coefficients; and $(\eta_{0j}, \eta_{1j})^{'}$ is the residual vector that follows $N(0, \Sigma)$, assumed to be independent of ϵ_{ij}. In the movie-level model, we set the intercept and the linear slope of the trajectory to vary among the movies in terms of both the effects of the movie-level covariates and random error. In particular, we assumed the coefficient of t_{ij} to be a function of online WOM variables, which determines the speed with which cumulative audience size grows.

Incorporating all possible covariates in the model may induce multicollinearity among them. Therefore, we employed the backward selection method to reduce the number of covariates in the above model. We used the statistical analysis software R to fit the model. Using the lme4 (Bates and Maechler, 2010) and lmerTest (Kuznetsova and Brockhoff, 2013) packages in R, we removed the least significant variables from the model sequentially until all covariates were significant.

Ⅳ. Results

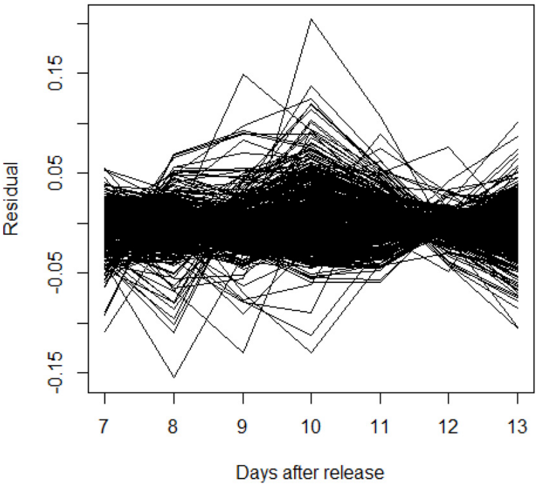
4.1. Parameter Estimates and Mean Trajectories

The final model selected from the backward selection procedure includes the linear, quadratic, and cubic term of t_{ij} (Time) and the covariates listed on <Table 3>. For the categorical predictors, the variables listed in <Table 3> represent the dummy variables for the category indicated in the parenthesis. For example, "Nationality (group2)" means the dummy variable having one for American movies (group 2.)

<Table 3> provides the fixed coefficient estimates and the p-values obtained via Satterthwaite approximation (1946), standard deviations of random errors, and correlations among them. The fitted day-level model is expressed as

$$\widehat{y_{ih}} = \widehat{\beta_{0j}} + \widehat{\beta_{1j}} t_{ij} - 0.016\, t_{ij}^{2} + 0.003\, t_{ij}^{3} - 0.107\,(Days\ of\ week)$$

and the fitted movie-level model is expressed as



<Figure 4> Residuals Plot

$$\begin{aligned}\widehat{\beta_{0j}} = &- 0.068 + 1.018\log \\ &(current.cum.num.of\,audience) \\ &- 0.028(Nationality: USA) \\ &- 0.024(Nationality: Others) \\ &+ 0.007(Month: group2) \\ &+ 0.015(Month: group3) \\ &+ 0.008(Score.after) \\ &- 0.004\log(Count.after + 1) \\ \widehat{\beta_{1j}} = &- 0.004 + 0.007(Score.after) \\ &+ 0.002\log(Count.after + 1)\end{aligned}$$

The negative value of -0.107 for the coefficient associated with the *day of week* variable indicates that, on average, the cumulative audience size is 10.1 % $(= 1 - \exp(-0.107) \times 100)$ lower on weekdays than on weekends. This difference seems somewhat smaller than expected. However, given that the mean *cumulative* audience size was 912,613 (see <Table 2>) and the mean *daily* audience was 64,287, a 10.1% (92,174 audience member) difference in cumulative audience size between weekdays and weekends does not seem small.

As expected, in the movie-level model, the linear coefficient $\widehat{\beta_{ij}}$ was positively related with online review scores and counts after the movie's release. Online review scores and counts at the end of the first week have a positive influence on the speed with which cumulative audience size increased in the following week. Because the response variable is log-transformed, the increment of growth speed varied as a function of current audience size. The

<Table 3> Parameter Estimates of the Final Model

Fixed Effects				
		Estimate	Std.	Pr(> |t|)
Day level	(Intercept)	-0.068	0.049	0.168
	Time	-0.004	0.008	0.573
	$(\text{Time})^2$	-0.016	0.000	0.000
	$(\text{Time})^3$	0.003	0.000	0.000
	Day of week (group2)	-0.107	0.003	0.000
Movie level	Log(current cum. num. of audience)	1.018	0.004	0.000
	Nationality (group2)	-0.028	0.007	0.000
	Nationality (group3)	-0.024	0.010	0.017
	Month (group2)	0.007	0.008	0.349
	Month (group3)	0.015	0.008	0.058
	Score.after	0.008	0.002	0.000
	Log (Count.after +1)	-0.004	0.004	0.318
	Score.after*time	0.007	0.001	0.000
	Log (Count.after +1)*time	0.002	0.001	0.053
Random Effects				
Groups		Std.Dev.	Corr	
Movie	Intercept	0.065		
	Time	0.031	-0.14	
Residual		0.106		

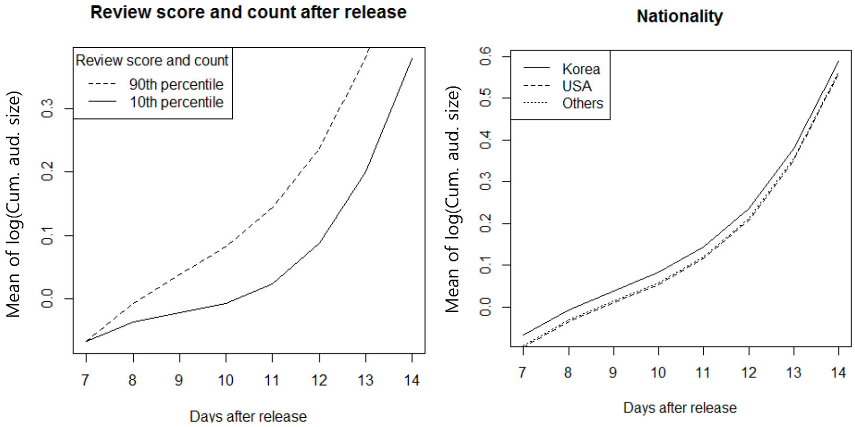
left panel of <Figure 5> includes the mean trajectories for the movies with review scores and counts of the 10th and 90th percentile in the sample. The left panel shows that the growth curve for films in the 10th percentile of online review scores and counts (at the end of the first week after release) increases much more slowly than films in the 90th percentile. This supports the claim that post-release online WOM influences the speed with which the cumulative audience size grows.

Because the intercept of the trajectory β_{0j} is expressed as a linear function of all movie-level covariates, their estimated coefficients will determine the overall level of the trajectory of the log-cumulative audience size. Current cumulative audience size, months other than July and August, and online review scores after a film's release exerted positive effects on future cumulative audience size. A film's internationality and the number of reviews it received after its release exerted negative effects on future cumulative audience size when the effects of the other covariates are adjusted. The right panel of <Figure 5> shows the mean trajectory of films that have been differentially categorized on the basis of their respective nationalities. Because nationality is included in the movie-level model only for β_{0j}, all nationality growth curves are parallel with each other. Korean movies have cumulative audiences that are 2.8% (=1-exp(-0.028)) and 2.4% (=1-exp(-0.024)) greater than American movies and movies from other countries, respectively.

4.2. Model Validation

<Figure 4> displays the residuals from the fitted model over time. We are generally unable to identify any systematic inflation of residual variance or curvature. Therefore, we infer that the model properly captures the trajectory pattern of the cumulative audience size.

To assess the result of the analyses presented in the previous sections, we used a 10-fold cross validation. Specifically, we randomly assigned all 777 movies into 10 groups such that the groups contained nearly equal numbers of movies. We fitted the model with a training set of nine movie groups, and tested the model with the remaining movie group. We then repeated this process 10 times, allowing a different



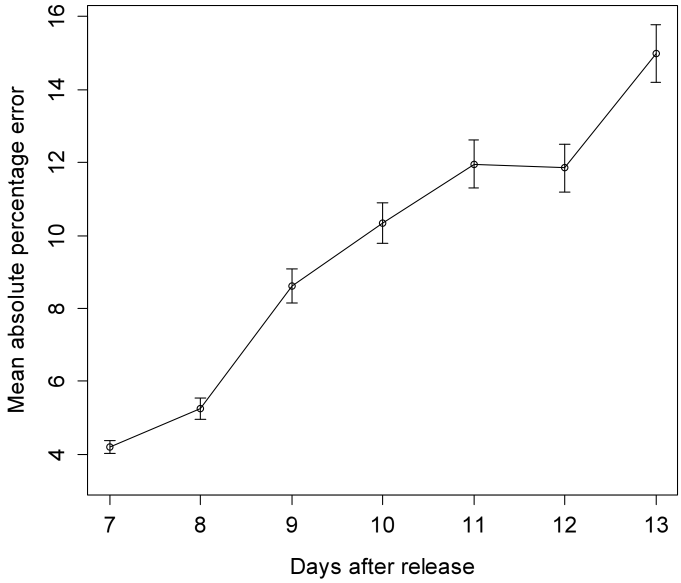
<Figure 5> Comparison of Mean Trajectories by Online Review and Nationality

group of movies to serve as the model test for each iteration. Subsequently, we plotted the mean absolute percentage error (MAPE) with the days following a film's release (see <Figure 6>). The MAPE is about 4% on the 7th day after a film's release and increases gradually to about 15% on the 13th day. It is natural that errors increase with the targeted prediction day being more distant from the current time. The average accuracy over the time is 90.4%, which seems acceptable.

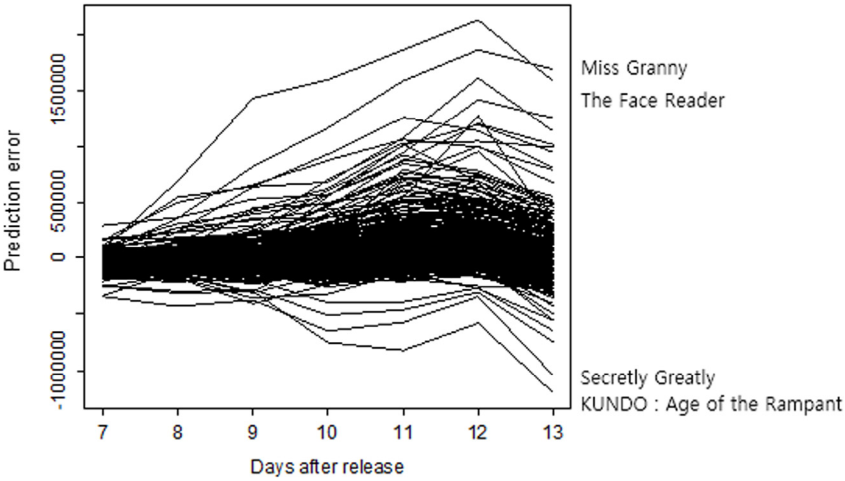
<Figure 7> illustrates the degree to which predictions for each movie were erroneous. *Miss Granny* and *The Face Reader* were characterized by large positive errors, indicating that they were more successful than the model predicted. *Miss Granny*, in particular, enjoyed significant box-office success, largely as a function of very positive online WOM (review score of 9) that seemed to increase its cumulative audience size at a greater rate than predicted. In contrast, *Secretly Greatly* and *Kundo* were less successful than model predicted. Both films garnered substantial public attention upon their initial release (both films had star casts), but the poor quality of the movies became known to audiences, thereby depressing their respective box-office successes.

Ⅴ. Conclusion

In this paper, we have developed and utilized a model for fitting the growth curves of cumulative audience sizes and forecasting the box-office performances of movies on specific future dates. The proposed model is particularly suited for exploring the effect of online WOM on the shape of the growth curve. Based on the hierarchical linear model, we fitted the mean trajectory of the cumulative audience size as a cubic function of time, and expressed the linear slope as a function of online WOM variables. Movies with superior online review scores or counts



<Figure 6> Mean Absolute Percentage Error of Prediction



<Figure 7> Prediction Errors

following their releases tend to have steep growth curves. Overall, the mean trajectory of a film's box-office performance was larger when the film was domestic, released in the summer, enjoyed positive online reviews after its release, or had significant attendance in the first week after its release.

Although we set only the intercept and the linear slope to be explained by the movie-level covariates, the quadratic and cubic slope can also be set to vary by movie, making the shape of the mean trajectory lines more flexible. However, this approach would increase the number of parameters to be estimated in the model and could cause problems with convergence during the optimization process. As a result, using more flexible models to capture the characteristics of each movie must be approached with caution. The proposed model can be combined with other prediction models (e.g., diffusion models, auto regressive models) that also capture the dynamic nature of box-office performance. We expect that an ensemble model merging these methods would increase the accuracy with which movie performance can be predicted.

<References>

[1] Ainslie, A., Dreze, X., and Zufryden, F. (2005). Modeling Movie Life Cycles and Market Share. *Marketing Science, 24*(3), 508-517.

[2] Asur, S., and Huberman, B. A. (2010). Predicting the Future with Social Media. In *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on.* IEEE, 1, 492-499.

[3] Bates, D., and Maechler, M. (2010). lme4: Linear Mixed-Effects Models Using S4 Classes. *R Package Version,* 0.999375-37.

[4] Basuroy, S., Chatterjee, S., and Ravid, S. A. (2003). How Critical Are Critical Reviews? The Box Office Effects of Film Critics, Star Power, and Budgets. *Journal of Marketing, 67*(4), 103-117.

[5] Chang, B. H., and Ki, E. J. (2005). Devising a Practical

Model for Predicting Theatrical Movie Success: Focusing on the Experience Good Property. *Journal of Media Economics*, *18*(4), 247-269.

[6] Chintagunta, P. K., Gopinath, S., and Venkataraman, S. (2010). The Effects of Online User Reviews on Movie Box Office Performance: Accounting for Sequential Rollout and Aggregation across Local Markets. *Marketing Science, 29*(5), 944-957.

[7] Dellarocas, C., Zhang, X. M., and Awad, N. F. (2007). Exploring the Value of Online Product Reviews in Forecasting Sales: The Case of Motion Pictures. *Journal of Interactive Marketing, 21*(4), 23-45.

[8] Draper, D. (1995). Assessment and Propagation of Model Uncertainty. *Journal of the Royal Statistical Society. Series B, 57*(1), 45-97.

[9] Duan, W., Gu, B., and Whinston, A. B. (2008). The Dynamics of Online Word-of-Mouth and Product Sales - An Empirical Investigation of the Movie Industry. *Journal of Retailing, 84*(2), 233-242.

[10] Eliashberg, J., and Shugan, S. M. (1997). Film Critics: Influencers or Predictors? *The Journal of Marketing, 61*(2), 68-78.

[11] Gelman, A., and Meng, X. (1996). Model Checking and Model Improvement. In *Markov Chain Monte Carlo in Practice* (pp. 189-201). London: Chapman and Hall.

[12] Jang, B. H., Lee, Y. H., and Nam, S. H. (2009). Elaborating Movie Performance Forecast through Psychological Variables: Focusing on the First Week Performance. *Korean Journal of Journalism & Communication Studies*, *53*(4), 346-371.

[13] Jedidi, K., Krider, R., and Weinberg, C. (1998). Clustering at the Movies. *Marketing Letters*, *9*(4), 393-405.

[14] Kim, J. H., and Lee, H. I., (2009). The Online Movie Ratings and Box Office Performances between Korea and U.S.: Hollywood Films in Korea during 2007~2008. *Film Studies* (42), 163-203.

[15] Kim, Y. H., and Hong, J. H. (2011). A Study for the Development of Motion Picture Box-Office Prediction Model. *Communications for Statistical Applications and Methods*, *18*(6), 859-869.

[16] Kim, Y.-H., and Hong, J.-H. (2013). A Study for the Drivers of Movie Box-Office Performance. *Korean Journal of Applied Statistics*, *26*(3), 441－452.

[17] Kim, Y. S., Im, S. H., and Jung, Y. S. (2010). A Comparison Study of the Determinants of Performance of Motion Pictures: A Comparison Study of the Determinants of Performance of Motion Pictures: Art Film vs. Commercial Film. *The Korea Contents Association*, *10*(2), 381-393.

[18] Kuznetsova, A., Brockhoff, P. B., and Christensen, R. H. B. (2013). lmerTest: Tests for Random and Fixed Effects for Linear Mixed Effect Models (lmer Objects of lme4 Package). *R Package Version*, 2-0.

[19] Lee, J. H., Park, J. S., Kim, H. M., and Park, J. H. (2013). Investigating the Influence of Perceived Usefulness and Self-Efficacy on Online WOM Adoption Based on Cognitive Dissonance Theory: Stick to Your Own Preference vs. Follow What Others Said. *Asia Pacific Journal of Information Systems*, *23*(3), 131-154.

[20] Lee, K. J., and Chang, W. J. (2006). Predicting Financial Success of a Movie Using Bayesian Choice Model. *Proceeding of the Korean Industrial Engineers Conference*, 1428-1433

[21] Liu, Y. (2006). Word of Mouth for Movies: Its Dynamics and Impact on Box Office Revenue. *Journal of Marketing, 70*(3), 74-89.

[22] Litman, B. R., and Kohl, L. S. (1989). Predicting Financial Success of Motion Pictures: The '80s Experience. *Journal of Media Economics*, *2*(2), 35-50.

[23] Park, S. H., Song, H. J., and Jung, W. K. (2011). The Determinants of Motion Picture Box Office Performance: Evidence from Korean Movies Released in 2009-2010. *Korea Regional Communication Research Association, 11*(4), 231-258.

[24] Park, S. H., and Song, H. J. (2012). Word of Mouth and Box Office Performance: WOM's Impact on Weekly Box Office Revenues. *Korean Journal of Journalism & Communication Studies, 56*(4), 210-235.

[25] Ravid, S. A. (1999). Information, Blockbusters, and Stars: A Study of the Film Industry. *Journal of Business, 72*(4), 463-492.

[26] Rui, H., Liu, Y., and Whinston, A. (2013). Whose and What Chatter Matters? The Effect of Tweets on Movie sales. *Decision Support Systems, 55*(4), 863-870.

[27] Satterthwaite, F. E. (1946). An Approximate Distribution of Estimates of Variance Components. *Biometrics Bulletin*, 110-114.

[28] Sharda, R., and Delen, D. (2006). Predicting Box-Office Success of Motion Pictures with Neural Networks. *Expert Systems with Applications*, *30*(2), 243-254.

[29] Sochay, S. (1994). Predicting the Performance of Motion Pictures. *Journal of Media Economics*, *7*(4), 1-20.

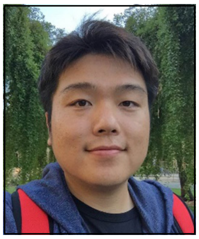
[30] Swami, S., Eliashberg, J., and Weinberg, C. B. (1999). Silver Screener: A Modeling Approach to Movie Screens Management. *Marketing Science, 18*(3), 352-372.

[31] Sawhney, M. S., and Eliashberg, J. (1996). A Parsimonious Model for Forecasting Gross Box-Office Revenues of Motion Pictures. *Marketing Science, 15*(2), 113-131.

[32] Wyatt, J. (1991). High Concept, Product Differentiation, and the Contemporary US Film Industry. *Current Research in Film: Audiences, Economics and Law, 5,* 86-105.

[33] Zhang L, Luo J, Yang S. (2009) Forecasting Box Office Revenue of Movies with BP Neural Network. *Expert Systems With Applications*, *36*, 6580-6587.

◆ About the Authors ◆



Jongmin Park

Jongmin Park received his Master degree in Business Administration from Kookmin University, Korea in 2015. He worked as a researcher at the Research21 and the Korea Research in Seoul, Korea, from 2010 to 2013. His research interests include analyzing of complex information, marketing analytics, and data mining methods.

Yeojin Chung

Yeojin Chung is an assistant professor of School of Business Administration at Kookmin University in Seoul, Korea. She worked as a postdoctoral researcher at the Graduate School of Education in University of California, Berkeley, from 2010 to 2013. She earned her Ph.D. degree in Statistics from the Pennsylvania State University at University Park in 2010. Her research interests include statistical methods for recommendation system, estimation problem in hierarchical linear models, clustering method based on density estimation, and nonparametric maximum likelihood methods.

Yoonho Cho

Yoonho Cho received the B.S. degree in computer science & statistics from Seoul National University and the Ph.D. degree in management engineering from the Korea Advanced Institute of Science and Technology (KAIST). He is a professor of School of Business Administration at Kookmin University. His research interests include business analytics, big data mining, recommender systems, and social network analysis.

Submitted: December 31, 2014; 1st Revision: May 14, 2015; Accepted: July 10, 2015