The Korean Journal of Applied Statistics (2018) DOI: http://dx.doi.org/10.5351/KJAS.2018.31.4.447
31(4), 447–461

Prediction of movie audience numbers using hybrid model combining GLS and Bass models

Bokyung Kim[a] · Changwon Lim[a,1]

[a]Department of Applied Statistics, Chung-Ang University

(Received March 5, 2018; Revised May 24, 2018; Accepted June 20, 2018)

Abstract

Domestic film industry sales are increasing every year. Theaters are the primary sales channels for movies and the number of audiences using the theater affects additional selling rights. Therefore, the number of audiences using the theater is an important factor directly linked to movie industry sales. In this paper we consider a hybrid model that combines a multiple linear regression model and the Bass model to predict the audience numbers for a specific day. By combining the two models, the predictive value of the regression analysis was corrected to that of the Bass model. In the analysis, three films with different release dates were used. All subset regression method is used to generate all possible combinations and 5-fold cross validation to estimate the model 5 times. In this case, the predicted value is obtained from the model with the smallest root mean square error and then combined with the predicted value of the Bass model to obtain the final predicted value. With the existence of past data, it was confirmed that the weight of the Bass model increases and the compensation is added to the predicted value.

Keywords: Korean Film Council, generalized least squares method, multiple regression, Bass model, hybrid model

1. 서론

영화진흥위원회의 한국 영화산업결산에 따르면 한국 영화 시장은 성숙 단계에 접어들었다는 평가를 받고 있고, 인구 1인당 연 평균 관람 횟수는 4.2회로 세계에서 가장 높은 수준을 보이고 있다. 국내 영화 산업 매출 또한 매년 꾸준히 증가하고 있다. 영화 산업 매출은 크게 극장, 디지털 온라인 시장, 해외 매출 세 가지로 나뉜다. 극장 매출은 영화 산업 매출 중 약 80% 정도로 가장 큰 비중을 차지한다. 극장은 1차 판매 경로이며, 극장을 이용하는 관객 수는 부가판권에 영향을 주기 때문에 중요한 영화 유통 채널이다. 이처럼 영화의 관객 수는 영화 산업 매출에 직결되는 중요한 요소이다 (Korean Film Council, 2016).

이에 따라 다양한 예측 모형을 통해 영화의 관객 수나 순수익으로 영화의 흥행 여부를 예측하는 연구가 활발히 이루어지고 있다. Lee와 Jang (2006)은 추정 모수의 사전분포를 모호사전분포로 정의함으로써

This research was supported by Next-Generation Information Computing Development Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT (No. 2017M3C4 A7083281).

[1]Corresponding author: Department of Applied Statistics, Chung-Ang University, 84 Heukseok-ro, Dongjak-gu, Seoul 06974, Korea. E-mail: clim@cau.ac.kr

변수들의 불확실성을 반영할 수 있고, 영화 간의 이질성을 고려할 수 있는 베이지안 선택 모형을 제안하였다. 베이지안 선택 모형으로 영화 흥행을 결정하는 요인을 탐색하여 영화 흥행 성과를 예측하였고 인공신경망과 비교한 결과, 베이지안 선택 방법이 상업적으로 성공한 영화를 예측하는 데에 더 좋은 성능을 보였다. Song과 Han (2013)은 선형모형(linear regression), 랜덤포레스트(random forest), gradient boosting model을 이용하여 영화의 순수익을 예측하였다. 한국 박스오피스 시장에서 순수익이 5억 원 이상인 영화를 대상으로 진행되었으며, 순수익을 백분위수에 따라 10개 범주로 나누어 각 범주에 1점부터 10점까지 준 점수를 양적 종속변수로 사용하여 예측하였다. 독립변수로는 영화장르, 등급, 속편 여부, 감독, 배우, 명절 개봉 일 여부, 방학 개봉 일 여부, 개봉 월, 개봉 월 평균 기온, 국내 영화의 참여 비율 등과 같은 영화 속성변수만을 사용하였다. 선형모형, 랜덤포레스트, gradient boosting model의 세 가지 예측 모형을 비교한 결과 gradient boosting model이 가장 우수하였다. 또한 Jung과 Yang (2013)은 여러 영화 흥행 요인들을 다중회귀분석을 통해 중요요인들을 결정한 뒤 흥행 영화를 예측하였다. 결정된 요인들을 통해서 다중회귀분석(multiple regression analysis), 의사결정나무(decision tree), 인공신경망(artificial neural network; ANN)의 예측 모형을 이용해 예측하였다. 그 결과 유의한 독립변수만 사용한 예측 모형의 정확도가 모든 독립변수를 사용했을 때보다 평균 8.2% 향상되었다.

이렇게 영화의 속성을 나타내는 변수를 사용해 흥행 여부를 예측하는 것 이외에 구전효과(word of mouth; WOM)를 기반으로 관객수를 예측하는 연구도 이루어지고 있다. Park 등 (2015)은 위계선형모형(hierarchical linear model)을 기반으로 하여 영화의 관객수 성장 곡선을 예측함으로써 관객수를 예측하였다. 이때 온라인 구전효과(online word of mouth)의 영향도 고려하였는데, 개봉 이후 높은 온라인 리뷰 점수를 가진 영화는 가파른 성장 곡선을 나타내는 것을 확인하였다. Jung 등 (2016)은 트위터 버즈량을 기반으로 한 Bass 확산 모형을 이용하여 영화의 수요 확산 패턴을 분석하였다. 기존 흥행 결정요인인 스크린수와 트위터 버즈량(구전효과)을 비교하여, 어떤 것이 실제 관객 트렌드를 정확하게 반영하는지를 Bass 확산모델을 기반하여 분석하였다. 그 결과, 트위터 버즈량 기반의 Bass 모형이 스크린수보다 실제 관객 트렌드를 정확하게 반영함을 확인하였다. 그리고 Jeon과 Son (2016)은 영화의 속성 이외에도 평점과 평가자 수, 블로그 수, 뉴스 수와 같은 온라인 구전효과의 영향을 고려하였다. 이후 의사결정나무, 다항로짓모형, support vector machine과 같은 여러 데이터마이닝 분류 기법들을 사용하여 관객수를 예측하였다. 영화의 속성만을 변수로 사용했을 때보다 구전효과에 대한 변수를 추가함으로써 예측 정확도가 더 높아지는 것을 확인하였다.

다중회귀모형은 선형모형으로서 영화의 속성을 독립변수로 사용하기 때문에 속성이 영화의 관객 수에 어떤 영향을 끼치는지 설명이 가능한 모형이다. 반면 Bass 모형은 S자 형태의 곡선으로 수요를 예측하는 비선형모형이다. 본 논문에서는 다중회귀모형과 Bass 모형을 결합한 Hybrid 모형을 통해 영화의 관객 수를 예측하고자 한다. Hybrid 모형에 관한 연구로 Zhang (2003)의 시계열 예측을 위해 ARIMA-신경망 모델을 결합한 연구가 있다. ARIMA의 추정치와 신경망의 추정치의 합으로써 Hybrid 모형의 예측치를 도출하였는데, 이는 선형예측기법과 비선형예측기법을 접목시키기 위한 방법이다. Wolf의 흑점자료와 환율자료를 통해 예측력을 비교한 결과 Hybird 모형의 예측력이 우수한 것으로 나타났다.

본 논문에서는 이러한 Hybrid 모형을 이용하여 베리 굿 걸, 제보자, 드라큘라 세 영화를 타겟으로 특정일의 누적관객수를 예측하였다. 예측이 이루어진 시점은 14년 10월 5일이며 각각의 영화에 대해 개봉 후 17일차, 10일차, 4일차인 14년 10월 12일의 누적관객수를 예측하였다. 전처리 및 모든 분석 과정은 통계프로그램인 R(R Core Team, 2014)의 여러 함수와 패키지를 이용하여 이루어졌다. 다중회귀모형에서는 변수선택법으로 All subset regression을 사용하였고 변수 선택 기준으로는 제곱근평균제곱오차(root mean square error; RMSE)를 사용하였다. 이 때 제곱근평균제곱오차는 5중 교차검증(5-fold cross validation) 방법을 통해 계산되었다.

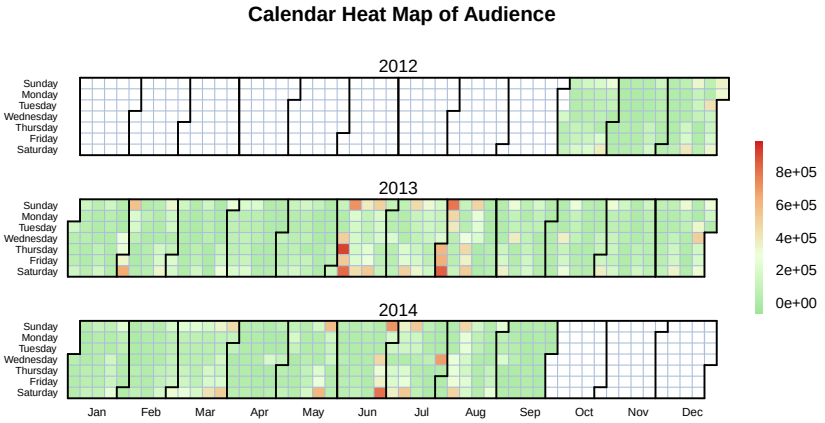


Figure 2.1. Heatmap of movie audience (2012.10.01–2014.10.05).

본 논문은 총 5장으로 구성되어 있다. 2장에서는 본 연구에서 사용한 데이터에 대한 설명과 이후 모델 수립 시 사용될 변수들에 대해 서술하였다. 3장에서는 예측에 사용된 일반화최소제곱법을 사용한 다중회귀모형, Bass 모형 그리고 이 둘을 결합한 Hybrid 모형에 대해 설명하였다. 4장에서는 Hybrid 모형을 이용한 베리 굿 걸, 제보자, 드라큘라 세 가지 영화의 예측 결과를 제시하였다. 마지막으로 5장에서는 본 논문을 종합적으로 정리하였고 추후 연구 방향에 대하여 논하였다.

2. 분석 데이터 설명

2.1. 데이터 수집

본 연구에서 사용된 데이터는 영화진흥위원회(www.kofic.or.kr)의 제공 데이터인 일별 박스오피스 데이터와 다음 무비(www.movie.daum.net)의 댓글 데이터이다. 계절성 효과의 패턴을 확인하기 위해 일별 박스오피스 데이터는 2012년 10월 1일부터 2014년 10월 5일까지, 약 2년 치의 데이터를 수집하였다. Figure 2.1은 박스오피스 데이터의 2년간 누적관객수를 나타낸 히트맵이다. 박스오피스 데이터에는 해당 일자에 상영 중인 모든 영화에 대해 개봉일, 매출액, 관객 수, 누적 관객 수, 스크린 수, 상영 횟수, 국적, 배우, 배급사 등 영화와 관련된 모든 정보가 포함되어 있다. 댓글 데이터는 다음 무비의 네티즌 평점 URL을 이용해 개봉 전과 개봉 후 각각의 데이터를 페이지 단위로 크롤링(Crawling)하여 영화 제목, 평점, 댓글, 작성일의 정보를 수집하였다. 댓글데이터도 박스오피스 수집 기간과 마찬가지로 작성일을 기준으로 2012년 10월 1일부터 2014년 10월 5일까지의 데이터를 수집하였다. 개봉 전 작성된 댓글의 경우에는 개봉 일주일 전까지의 데이터만을 추출하였다.

영화진흥위원회는 상영날짜를 기준으로 일별 박스오피스 데이터를 제공하고 있다. 수집한 데이터의 기간은 2012년 10월 1일부터 2014년 10월 5일까지지만 이 중에는 2012년 10월 이전에 개봉된 영화나 재개봉된 영화가 다수 포함되어 있다. 따라서 개봉일을 기준으로 2012년 10월 1일 이후의 데이터만을 추출하였다. Figure 2.2는 영화의 누적관객수(단위: 만 명)에 따른 밀도함수그래프이며 그래프에 표시된 직선은 전체 누적관객수의 평균(약 11만 5천명)을 나타낸다. 누적관객수가 10만 미만인 영화의 개수는 전체 영화의 약 86%를 차지한다. Figure 2.3은 누적관객수가 10만 이상인 영화의 매출액 합계와 10만 미만인 영화의 매출액 합계를 비교하기 위한 그림이다. 누적관객수가 10만 이상인 영화의 매출액은 전체 매출액의 98%인 반면, 누적관객수가 10만 미만인 영화의 매출액은 2%밖에 되지 않는다. 따라서 그

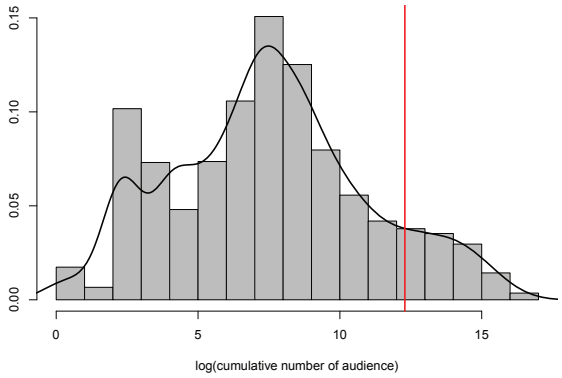


Figure 2.2. Histogram of cumulative number of audience.

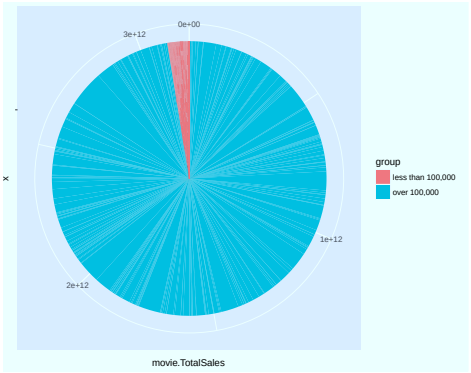


Figure 2.3. Comparison of total sales.

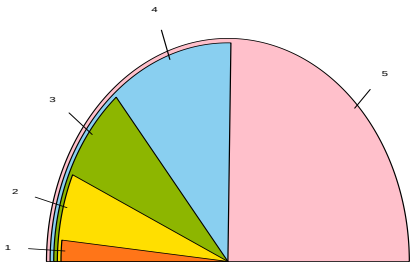


Figure 2.4. Fanplot of director group.

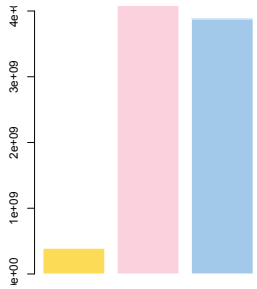


Figure 2.5. Barplot of distributor group.

갯수에 비해 매출액이 미미하기 때문에 누적관객수가 10만 명 미만인 영화도 제외하였다. 최종적으로 분석에는 영화 268편이 사용되었다. 타겟 영화인 베리 굿 걸, 제보자, 드라큘라의 개봉 후 17일차, 10일차, 3일차 누적관객수를 예측하는 것이 목적이기 때문에 개봉 후 17일차까지의 데이터만을 추출하였다.

2.2. 변수 설명

본 연구는 여러 요인들에 대한 정보들을 가지고 영화의 누적 관객 수를 예측하는 것이다. 반응변수인 누적 관객 수를 잘 설명할 수 있는 독립변수를 설정하는 것이 중요하다. Park 등 (2015)에 따라 요인이 다양한 감독, 배급사, 요일, 월, 등급, 장르는 군집으로 축소하였다. 분석에 사용된 독립변수는 총 21개이며 아래에서 자세히 설명하고자 한다.

① 감독군, 배급사군, 장르군

감독, 배급사, 장르 변수가 2개 이상의 값을 가지는 경우에는 맨 앞에 있는 값을 대푯값으로 설정하여 K-means 군집분석을 통해 군집으로 축소하였다. 감독평균최대관객수는 감독별 누적관객수의 평균을 의미하며 이 값이 높은 순으로 감독 그룹$(5, 4, 3, 2, 1)$을 설정하였고(Figure 2.4), 좀 더 세세한 반영을 위해 감독평균최대관객수 또한 변수로 추가하였다. 배급사의 경우에는 배급사평균최대관객수에 따라 3가지 군집으로 축소한 뒤(Figure 2.5) 배급사평균최대관객수를 변수로 추가하였고, 장르 또한 장르평균최대관객수에 따라 4개의 군집으로 축소한 뒤(Figure 2.6) 장르평균최대관객수도 함께 변수로 추가하였다.

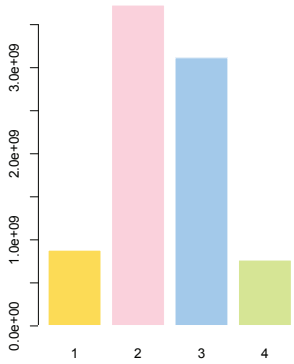


Figure 2.6. Barplot of genre group.

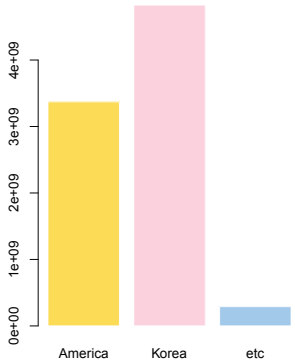


Figure 2.7. Barplot of nationality group.

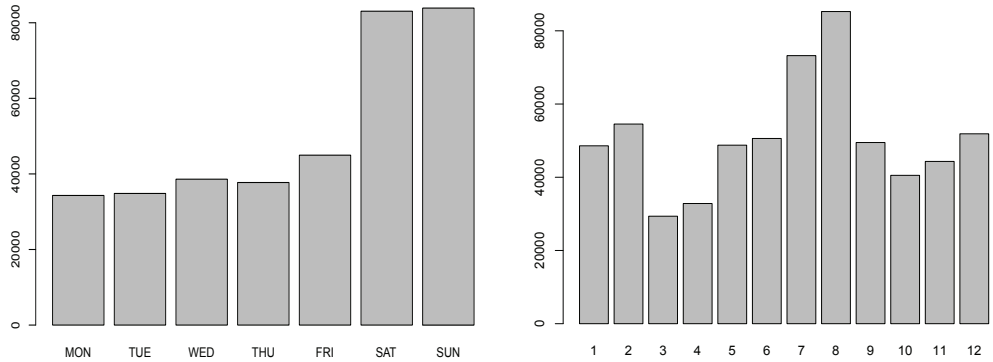


Figure 2.8. Barplot of average audience by day and month.

② 국적군

각 영화의 대표 국적으로 그룹을 나누었으며 한국, 미국 그리고 기타 세 개의 그룹으로 나누었다. Figure 2.7은 국적군별 누적관객수 그래프이며 한국과 미국 그리고 나머지 그룹의 누적관객수가 상이한 것을 확인할 수 있다.

③ 요일군, 월군

요일군과 월군은 요일 또는 월별로 평균 누적관객수의 양상이 비슷한 것을 묶은 변수이다. Figure 2.8을 보면 요일별로 평균 누적관객수의 양상이 상이한 것을 확인할 수 있으며 평균 누적관객수가 비슷한 것끼리 묶어 월, 화, 수, 목과 토, 일 그리고 금요일, 세 가지의 그룹으로 나누었다. 월도 마찬가지로 평균 누적관객수가 비슷한 것끼리 묶어 7, 8월과 1, 2, 5, 6, 9, 12월 그리고 3, 4, 10, 11월의 세 가지 그룹으로 나누었다.

④ 등급군

관람 등급에는 15세 관람가, 12세 미만인 자는 관람할 수 없는 등급, 고등학생 이상 관람가 등 다양한 등급으로 나눠져 있어 4가지 등급(전체관람가, 12세 이상 관람가, 15세 이상 관람가, 청소년 관람불가)으로 간소화하였다. Figure 2.9는 등급별로 누적관객수를 나타낸 그래프이다. 그래프에서 네모칸의 크기는 스크린 수를, 네모칸의 색깔은 누적관객수를 나타낸다. 15세 이상 관람가 영화의 스크린 수와 누적관

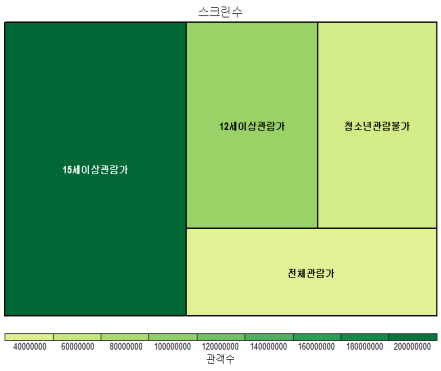


Figure 2.9. Treemap of ranking group.

Table 2.1. Description of variables

Variables	Description
Input variables	
log(현시점누적관객수)	예측하는 시점(12년 10월 5일)의 누적관객수
일차	개봉 후 경과 일수 (각 날짜 ? 개봉일)
일차2	'일차' 변수값의 제곱값
일차3	'일차' 변수값의 세제곱값
log(감독평균최대관객수)	해당 감독 영화의 누적관객수 평균
감독군	감독별 평균 누적관객수에 따라 감독을 5개 그룹으로 간소화
log(배급사평균최대관객수)	해당 배급사 영화의 누적관객수 평균
배급사군	배급사별 평균 누적관객수에 따라 배급사를 3개 그룹으로 간소화
요일군	요일별 평균 누적관객수에 따라 3개 그룹으로 간소화
등급군	전체관람가, 12세 이상, 15세 이상, 청소년관람불가 4개 등급으로 간소화
국적군	한국, 미국, 기타 3개의 그룹으로 간소화
log(장르평균최대관객수)	해당 장르 영화의 누적관객수 평균
장르군	장르별 평균 누적관객수에 따라 4개 그룹으로 간소화
월군	월별 평균 누적관객수에 따라 월을 3개 그룹으로 간소화
월9/월10	타겟 영화의 개봉 월을 더미변수로 추가
점유율군	해당 일차에 해당하는 지정 영화의 점유율을 3개 그룹으로 간소화
n일차	개봉일로부터 해당영화의 상영일차까지의 누적관객수
log(before.댓글수 + 1)	개봉 일주일전까지의 해당영화 댓글 수
log(before.평점평균 + 1)	개봉 일주일전까지의 해당영화 평점평균
log(after.댓글수 + 1)	개봉 후 해당영화의 댓글 수
log(after.평점평균 + 1)	개봉 후 해당영화의 평점평균
Response variables	
log(누적관객수)	

객수가 가장 많으며 등급에 따라 누적관객수가 상이한 것을 확인할 수 있다.

⑤ 월9, 월10

타겟 영화인 베리 굿 걸, 제보자의 개봉월은 9월이고 드라큘라의 개봉월은 10월이다. 예측에 좀 더 세밀하게 영향을 주고자 9월과 10월을 더미변수로 추가하였다.

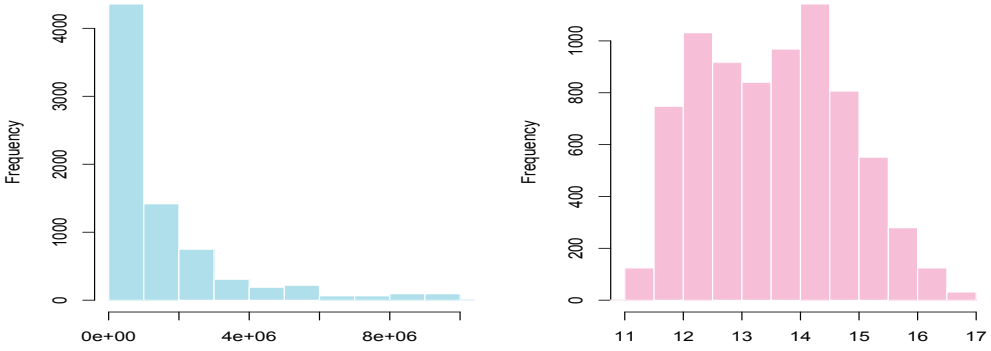


Figure 2.10. Distribution of director average maximum number of audiences before/after log-transformation.

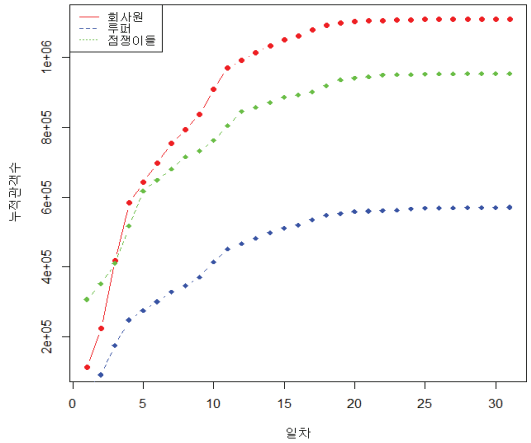


Figure 2.11. Cumulative moive audience by day release.

⑥ n일차

개봉 일차에 따른 누적관객수는 일정 패턴을 보유하고 있기 때문에 개봉일로부터 n일차의 상영일까지 해당 영화의 누적관객수를 계산하여 일차 컬럼으로 할당한다. 베리 굿 걸의 경우 개봉 후 17일차, 제보자의 경우 개봉 후 10일차 그리고 드라큘라의 경우 개봉 후 4일차를 예측하기 때문에 0일차부터 17일차까지, 총 17개 컬럼을 할당한다.

⑦ Before/After 댓글 수와 평점평균

다음 무비의 네티즌 평점 URL을 이용해 개봉 전과 개봉 후의 댓글 데이터 각각을 페이지 단위로 크롤링하였다. 수집 기간은 박스오피스 데이터와 마찬가지로 2012년 10월 1일부터 2014년 10월 5일까지이며 개봉 전과 개봉 후 댓글 수와 평점 평균을 변수로 추가하였다.

Table 2.1은 분석에 사용한 전체 변수들을 요약한 표이다. 값의 분포를 고르게 하기 위해 현시점누적관객수, 감독평균최대관객수, 배급사평균최대관객수, 장르평균최대관객수, 댓글 수, 평점평균은 로그변환을 하였다. Figure 2.10을 보면 로그변환 후에 분포가 고르게 나타나는 것을 확인할 수 있다. 댓글 수와 평점평균의 경우 0인 값도 존재하기 때문에 로그변환이 가능하도록 1을 더한 뒤에 로그변환을 사용하였다. Figure 2.11은 '회사원', '루퍼', '점쟁이들' 세 영화의 일차에 따른 누적관객수를 표현한 그림이다.

일차에 따른 누적관객수는 선형관계가 아닌 비선형 관계를 가지는 것을 확인할 수 있다. 이러한 비선형 효과를 고려하기 위해 일차변수는 제곱과 세제곱 변환을 하여 변수로 추가하였다.

3. 모델 수립

3.1. 일반화 최소제곱법을 이용한 다중회귀

다중회귀모형은 하나의 종속변수를 예측하기 위해 여러 독립변수들을 사용할 때 가장 흔히 사용되는 모형이다. 일반적인 다중회귀모형의 기본식은 다음과 같다.

$$y_{ij} = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij} + \cdots + \beta_p x_{pij} + \epsilon_{ij}. \tag{3.1}$$

본 연구는 영화 268편을 대상으로 수행되었다. 여기서 y_{ij}는 $i\,(i = 1, \ldots, 268)$번째 영화의 개봉 $j\,(j = 1, \ldots, 17)$일차의 log(누적관객수)이고 x_{1ij}부터 x_{pij}는 $p\,(p = 1, \ldots, 21)$개의 독립변수들의 i번째 영화의 개봉 j일차의 값이다. 다중회귀 모형의 기본 가정은 오차항이 서로 독립이고 분산이 일정한 정규분포를 따른다는 것이다. 하지만 본 연구의 데이터는 영화의 일차별 관객 수이며 같은 영화의 일차별 관객수는 서로 의존적이다. 가령 한 영화의 1일차 누적 관객 수와 2일차 누적관객수 사이에는 상관성이 있을 것이 자명하다. 따라서 오차항이 서로 독립이라는 가정은 성립하지 않는다.

이러한 오차항에 대한 가정을 완화시키기 위한 방법으로 오차항이 자기회귀(autoregressive; AR)모형을 따르는 일반화 최소제곱법(generalized least square; GLS)을 사용하였다. AR 모형은 아래의 수식과 같이 시점 j의 잔차가 $(j-1)$시점 혹은 그 이전의 잔차에 영향을 받는다고 가정한다. 본 연구에서는 j일차의 영화 관객수는 $(j-1)$일차의 관객수에 영향을 받을 것이라 가정하고 잔차가 AR(1)을 따르는 GLS 모형을 사용한다.

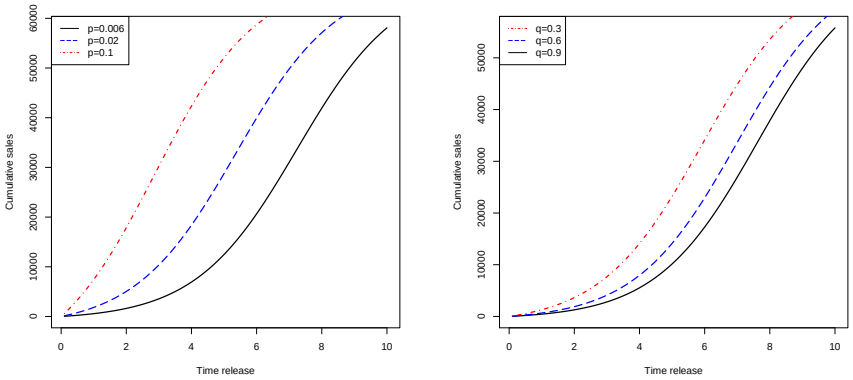
$$\epsilon_{ij} = \phi_1 \epsilon_{i(j-1)} + \upsilon_{ij}, \quad \upsilon_{ij} \overset{\text{iid}}{\sim} \left(0, \sigma_\upsilon^2\right). \tag{3.2}$$

다중회귀모형은 앞서 설명한 것처럼 여러 독립변수들을 사용해 하나의 종속변수를 예측한다. 영화의 누적관객수를 설명하는데 사용 가능한 21개의 독립변수를 모두 포함하여 모형을 수립하기 보다는 최적의 독립변수 조합을 찾아 모형을 수립하는 것이 적절하다고 판단하였다. 이를 위해 본 연구에서는 All-subset regression 방법을 사용한다. All-subset regression 방법은 모든 독립변수 조합으로 모형을 구축하고 사전에 정한 척도에 의해 가장 좋은 모형을 선택하는 방법이다. 반드시 필요하다고 판단되는 일차, 일차2, 일차3, 요일군, 월9나 월10, 현시점누적관객수의 7가지 변수를 고정하고 나머지 14개의 독립변수를 이용해 가능한 모든 조합인 2^{14}개, 총 16,384개 변수 조합으로 분석을 진행했다.

모든 독립변수 조합으로 이루어진 모형들 중 가장 좋은 모형을 선택하는 기준으로 본 논문에서는 제곱근평균오차(root mean-squared error; RMSE)를 사용하였고 제곱근평균오차를 계산하기 위해 5중 교차검증 방법을 사용하였다.

$$\text{RMSE} = \sqrt{\frac{1}{268 \times 17} \sum_{i=1}^{268} \sum_{j=1}^{17} \left(Y_{ij} - \widehat{Y}_{ij}\right)^2}, \tag{3.3}$$

여기서 Y_{ij}는 실제 i번째 영화의 j일차 누적관객수이며 $\widehat{Y}_{ij}$ 예측된 i번째 영화의 j일차 누적관객수이다. 5중 교차 검증 방법은 모델 수립을 위해 사용한 전체 데이터를 랜덤하게 5개의 그룹으로 나눈 뒤 4개의 그룹을 모형의 훈련 데이터(training data)로 사용하고 나머지 1개의 그룹을 검증용 데이터(test data)로 사용하여 모형의 정확도를 측정하는 방법이다. 검증용 데이터를 바꿔가며 독립적으로 5회 수행한 뒤 각 회차에서 얻은 5개의 정확도를 평균하여 모형의 정확도로 이용한다.



(a) Fixed q (b) Fixed p

Figure 3.1. Shape of Bass curve according to parameters.

시계열 자료의 특성을 반영하기 위해 데이터를 나눌 때 개봉 월이 5개의 그룹에 골고루 분배되도록 나눈다. 전체 영화의 개수는 268개이며 이를 각각 53개씩 5개의 그룹으로 나누고 한 그룹을 검증용 데이터로 나머지 4개의 그룹을 훈련용 데이터로 하여 모형을 추정한다. 이를 1만 6천여개의 모든 변수 조합에 대해 반복 시행하고 제곱근평균오차의 값이 최소가 되는 모형을 선택한다.

3.2. BASS 모형

Bass 모형은 상품이나 서비스의 수요량을 추정하는 확산 모형으로서 수요량의 패턴을 S자 형태의 곡선으로 추정하는 모형이다 (Bass, 1969). 과거의 자료가 축적되어 있으면 정확하게 수요를 산출할 가능성이 높아지며 수요 자료가 없는 신제품의 수요까지도 예측할 수 있다. 시간에 따른 수요량 자료만으로 모형을 추정하게 된다. 즉, 수요예측 기법 중 하나로써 과거 수요를 기초로 한 제품군의 미래 수요를 예측하는 데에 유용한 모형이다. 본 연구에 적용된 Bass 모형의 식은 다음과 같다.

$$Y_{ij} = m\frac{(p+q)^2}{p}\frac{e^{-(p+q)j}}{\left(1+\frac{q}{p}e^{-(p+q)j}\right)^2} + \epsilon_{ij}, \quad \epsilon_{ij} \overset{\text{iid}}{\sim} \left(0, \sigma^2\right). \tag{3.4}$$

다중회귀모형에서와 마찬가지로 여기서 Y_{ij}는 실제 i번째 영화의 j일차 누적관객수이며 회귀모수인 p는 혁신계수(innovation coefficient), q는 모방계수(imitation coefficient), m은 잠재시장 규모(potential number of ultimate adopters)를 나타낸다.

Bass 모형을 이용한 수요 예측은 혁신계수(p), 모방계수(q), 잠재시장 규모(m)의 세 가지 모수 추정을 필요로 한다. 본 연구에서는 일반적인 nonlinear least squares (NLS) 방법을 이용했다. Bass 모형은 기본적으로 선형이 아닌 S자 곡선의 형태를 지니는 특성을 가지기 때문에 NLS방법을 사용하며 이는 추정된 모형과 실제 데이터의 오차 제곱 합을 최소화하는 표준적인 비선형 모수 추정방법이다 (Venkatraman 등, 1994). 혁신계수(p)는 보통 0.0007에서 0.03 사이의 값을 가지고 모방계수(q)는 0.38에서 0.56 사이의 값을 가진다 (Talukdar 등, 2002). 혁신계수는 모방계수에 비해 확산에 많은 영향을 끼치지는 못하는 것으로 알려져 있고, 일반적인 확산 과정은 모방계수에 의해 주도된다는 특징이 있다. 이는 모방계수의 값이 클수록 확산이 빠르게 일어난다는 것을 의미한다. Figure 3.1은 확산계수와 모방계수 값의 변화에 따라 S자 곡선의 형태를 보여주는 그림이다. (a)는 모방계수를 고정시킨 뒤에 혁신계수를 변화

Table 4.1. Summary of final hybrid model by target movie

타겟 영화	예측시점 개봉 후 경과일	최종 독립변수	Bass 가중치
베리 굿 걸	0–10일	일차 + 일차2 + 일차3 + 요일군 + 월9 + log(현시점누적관객수) + log(감독평균최대관객수) + 장르군 + 점유율군 + log(before.댓글수 + 1)	0.3
제보자	0–3일	일차 + 일차2 + 일차3 + 요일군 + 월10 + log(현시점누적관객수) + log(감독평균최대관객수) + 장르군 + 점유율군 + 배급사군	0
드라큘라	미개봉	일차 + 일차2 + 일차3 + 요일군 + 월10 + log(감독평균최대관객수) + 등급군 + log(장르평균최대관객수) + 점유율군 + log(배급사평균최대관객수)	0

시켰을 때의 그래프로 혁신계수의 값이 증가할수록 그래프의 기울기가 급해지는 것을 볼 수 있다. (b)는 혁신계수를 고정시킨 뒤에 모방계수를 변화시켰을 때의 그래프로 모방계수의 값이 작아질수록 그래프의 기울기가 급해지는 것을 확인할 수 있다. 본 연구에서도 각 영화들에 Bass 모형을 적용한 뒤 모수들을 비교해봄으로써 영화들의 확산 정도를 측정해보고자 한다.

3.3. Hybrid 모형

본 연구에서 사용한 Hybrid 모형은 앞서 서술한 GLS를 이용한 회귀 모형과 Bass 모형을 결합한 형태이다. 모형의 결합은 GLS와 Bass 모형 각각의 예측치 합 (Zhang, 2003)이 아닌 두 모형에 적절한 가중치를 두어서 최소 RMSE를 내는 방법을 사용한다. 먼저 각 영화의 일차별 Bass 모형 예측치와 GLS 예측치를 도출한다. 이후 가중치를 0부터 1까지 0.1 간격으로 시행착오법을 적용하여 모든 영화에 대해 10개의 Hybrid 모형을 만들고 이 중 RMSE가 가장 작을 때의 가중치를 선택한다. Hybrid 모형을 사용하는 이유는 GLS 모형의 예측치를 Bass 모형의 예측치로 보정하기 위함이다. 실제 관객 수가 존재하는 개봉 영화의 경우 개봉 일차가 늘어남에 따라 Bass 모형의 보정효과가 올라갈 것으로 기대되며 미개봉 영화의 경우에는 과거의 자료가 존재하지 않기 때문에 Bass 모형의 가중치는 0이다.

$$\hat{y} = \omega \hat{y}_{\text{GLS}} + (1 - \omega)\hat{y}_{\text{Bass}} \tag{3.5}$$

여기에서, $\hat{y}_{\text{GLS}}$는 GLS 모형의 예측치이고, $\hat{y}_{\text{Bass}}$는 Bass 모형의 예측치이고, ω는 두 모형에 대한 가중치이다.

4. 분석 결과

본 연구의 예측 시점은 14년 10월 5일이며 타겟 영화들의 14년 10월 12일의 누적관객수를 예측하였다. 최종적으로 사용된 Hybrid 모형은 Table 4.1과 같다. '베리 굿 걸'의 경우는 10일치 데이터를 이용해 17일차의 누적관객수를 예측하였으며 이 때 선택된 독립변수로는 고정시킨 7개의 변수 이외에 감독평균최대관객수, 장르군, 점유율군, before.댓글수의 변수가 사용되었다. '제보자'의 경우는 3일치의 데이터를 이용해 10일차 누적관객수를 예측하였으며 선택된 독립변수로는 고정시킨 7개의 변수 이외에 감독평균최대관객수, 장르군, 점유율군, 배급사군의 변수가 사용되었다. '드라큘라'의 경우에는 미개봉 상태에서 4일차의 누적관객수를 예측하였고 독립변수로는 고정 변수 이외에 감독평균최대관객수, 등급군, 장르평균최대관객수, 점유율군, 배급사평균최대관객수의 변수가 사용되었다.

Table 4.2는 타겟 영화 3편을 예측하기 위해 사용된 GLS모형의 회귀계수, 표준오차, p-value를 정리한 표이다. 베리 굿 걸의 경우 월9, 감독평균최대관객수, 장르군이 유의하지 않은 결과가 나왔고, 제보자의 경우에는 요일군, 월10, 감독평균최대관객수, 장르군, 배급사군이 유의하지 않다는 결과가 나왔으

Table 4.2. The table of coefficients of GLS

	Variable	Coefficients	Std. error	p-value
베리 굿 걸	일차	0.6112	0.0050	0.0000
	일차2	−0.0511	0.0007	0.0000
	일차3	0.0014	0.00002	0.0000
	요일.group2	0.0074	0.0035	0.0342
	요일.group3	0.0772	0.0036	0.0000
	월9	−0.0536	0.0289	0.0636
	log(현시점누적관객수)	0.9715	0.0534	0.0000
	log(감독평균최대관객수)	−0.0243	0.0483	0.6146
	장르.group2	0.0692	0.0374	0.0641
	장르.group3	0.0851	0.0414	0.0400
	장르.group4	0.0647	0.1172	0.5807
	점유율.group2	0.0411	0.0082	0.0000
	점유율.group3	0.0384	0.0123	0.0018
	log(before.댓글수 + 1)	0.0339	0.0128	0.0080
제보자	일차	0.7491	0.0067	0.0000
	일차2	−0.0911	0.0014	0.0000
	일차3	0.0038	0.00008	0.0000
	요일.group2	−0.0031	0.0045	0.4902
	요일.group3	0.0982	0.0047	0.0000
	월10	0.0336	0.0281	0.2324
	log(현시점누적관객수)	0.7888	0.0369	0.0000
	log(감독평균최대관객수)	0.1941	0.0310	0.6146
	장르.group2	0.0423	0.0357	0.2361
	장르.group3	0.0162	0.0406	0.6902
	장르.group4	0.1565	0.1174	0.1825
	점유율.group2	0.0457	0.0100	0.0000
	점유율.group3	0.0377	0.0147	0.0103
	배급사.group2	−0.0326	0.0393	0.4073
	배급사.group3	−0.0264	0.0450	0.5572
드라큘라	일차	0.7991	0.0172	0.0000
	일차2	−0.0172	0.0078	0.0000
	일차3	0.0057	0.0010	0.0000
	요일.group2	−0.0587	0.0104	0.0000
	요일.group3	0.0910	0.0107	0.0000
	월10	0.0715	0.0547	0.1910
	log(감독평균최대관객수)	0.8046	0.0242	0.0000
	등급.group2	0.1805	0.0961	0.0605
	등급.group3	0.1260	0.0875	0.1499
	등급.group4	0.2905	0.0941	0.0021
	log(장르평균최대관객수)	0.1154	0.0744	0.1209
	점유율.group2	0.0744	0.0202	0.0002
	점유율.group3	0.0515	0.0307	0.0939
	log(배급사평균최대관객수)	0.1948	0.0363	0.0000

GLS = generalized least square.

Table 4.3. Parameter of Bass model

Parameter	베리 굿 걸			제보자		
	Estimate	p-value	Signif. code	Estimate	p-value	Signif. code
p	0.1637	0.000	***	0.00027	0.9994	
q	0.477	0.000	***	−0.2759	0.0394	*
M	94400	0.000	***	671400000	0.9994	

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1.

Table 4.4. The results of target movies audience predictions

타겟 영화	예측치	실제값	RMSE	예측률
베리 굿 걸	124,526	103,318	21,208	82.97%
제보자	1,227,764	1,258,835	31,071	97.53%
드라큘라	492,809	555,409	62,600	88.73%

RMSE = root mean square error.

Table 4.5. Comparison of predicted values based on models for 'Very Good Girl'

GLS	Bass	Hybrid	실제값
107,717	163,746	124,526	103,318

GLS = generalized least square.

며 드라큘라의 경우에는 월10, 등급군 등의 변수가 유의하지 않다는 결과가 나왔다. 이는 변수 선택시에 5-cross validation을 이용해 RMSE를 기준으로 변수를 선택했기 때문에 유의하지 않은 변수들이 포함된 것으로 생각되며 본 연구는 영화의 관객수를 예측하는 것이기 때문에 유의하지 않은 변수가 포함되어 있어도 무방하다. '베리 굿 걸'과 '제보자'에 사용된 변수들 중 일차변수와 현시점누적관객수 변수가 누적관객수에 가장 큰 영향을 미치고 있으며, 미개봉 상태였기 때문에 현시점누적관객수 변수가 없는 '드라큘라'의 경우에는 일차와 감독평균최대관객수가 누적관객수에 가장 큰 영향을 미치는 것을 확인할 수 있다.

Table 4.3은 Bass모형의 추정된 모수 p, q, M을 정리한 표이다. 베리 굿 걸에 사용된 Bass모형의 모수 추정치들은 모두 유의한 값을 가지고 있는 반면 제보자에 사용된 Bass모형의 모수 추정치들은 모방계수를 제외하고는 유의하지 않은 결과가 나왔다. 이는 Bass 모형의 특성상 과거데이터가 존재해야 하기 때문에 나타나는 현상이라고 볼 수 있다. 최종적으로 '베리 굿걸'은 Bass 모형을 결합한 Hybrid 모형이 적용되었지만 '제보자'나 '드라큘라'의 경우에는 Bass 모형이 결합되지 않은 GLS 모형이 적용되었다.

Table 4.4는 타겟 영화의 10월 12일 실제 누적관객수와 예측 누적관객수를 비교한 표이다. '제보자'가 예측률 97.5%로 가장 정확하게 예측하였고 '드라큘라'는 약 88.7%의 예측률을 보였다. 과거 데이터가 가장 많았던 '베리 굿 걸'은 약 83%로 비교적 저조한 예측률을 보였다. Table 4.5는 Hybrid 모형이 적용된 '베리 굿 걸'의 GLS, Bass, Hybrid 모형 예측값들을 비교한 표이다. Bass 모형이 실제값과 가장 큰 오차를 보이고 있고 GLS 모형이 실제값과 가장 근접한 결과를 나타냈다. 이러한 오차가 발생한 이유는 통상적으로 영화의 스크린 수는 2주 정도 유지되는 데에 반해 '베리 굿 걸'의 스크린 수는 개봉 일주일 만에 300에서 50으로 급감했다. 스크린 수의 급감이 오차의 원인이 된 것으로 판단된다. 또한 '드라큘라'는 감독의 데뷔작이기 때문에 감독에 대한 정보가 없었다. 감독군을 비슷한 장르의 감독평균최대관객수를 이용하여 설정하였는데 이것이 오차의 원인이라고 판단하였다.

'제보자'는 과거 데이터가 존재함에도 Bass모형의 가중치가 0이었기 때문에 Bass모형의 효용성을 확인하기 위해 추가로 13일차까지의 데이터를 수집해서 다시 분석해보았다. Table 4.6는 3일차까지의 데

Table 4.6. The results of an additional analysis for 'The Informant'

개봉 후 경과일	예측일자	Bass 가중치	Hybrid 예측값	실제값	RMSE
0–3일	10일차	0	1,227,764	1,258,835	31071
0–10일	20일차	0.2	1,623,978	1,626,847	2869

RMSE = root mean square error.

이터로 10일차를 예측한 경우와 10일차까지의 데이터로 20일차를 예측한 결과를 비교한 표이다. 기존 분석에서는 Bass 모형의 가중치가 0, RMSE가 31071이었다. 13일차까지의 데이터를 업데이트 한 뒤 20일차의 관객수를 예측해보았더니 Bass 모형의 가중치가 0.2, RMSE가 2869로 현저히 줄어든 것을 확인할 수 있었다.

5. 결론

본 연구에서는 영화의 특정 일차 누적관객수 예측을 위해 영화진흥위원회에서 제공하는 일별 박스오피스와 댓글 정보를 이용하여 여러 독립변수들을 생성하였다. GLS 방법을 이용한 회귀식을 결정을 위해 변수를 선택하는 방법으로는 All Subset regression 방법을 사용하였고 5중 교차 검증 방법을 통해 계산된 RMSE로 모형의 정확도를 평가하였다. 이후 Bass 모형과 가중 결합한 Hybrid 모형을 통해 관객 수를 예측하였다.

'베리 굿 걸'의 경우 과거 10일치의 데이터가 존재했고 '제보자'의 경우에는 과거 3일치의 데이터만 존재했다. '드라큘라'는 개봉이 되지 않은 상태였기 때문에 과거 데이터가 존재하지 않았다. 과거 데이터가 존재하는 베리 굿 걸, 제보자에서 Bass 모형의 효용이 나타날 것으로 기대했지만 베리 굿 걸에만 효용을 나타냈다. 제보자는 3일치 과거 데이터가 존재했음에도 Bass 모형의 가중치가 0이었다. 분석 이후 13일차까지의 데이터를 추가로 수집해서 '제보자'를 다시 분석해본 결과, Bass 모형의 가중치가 0.2로 증가하였고 RMSE는 2869로 현저히 줄어든 것을 확인할 수 있었다.

본 논문은 영화의 누적관객수 예측을 위해 GLS 모형과 Bass 모형을 결합한 Hybrid 모형을 적용해보았다. Hybrid 모형을 제안한 이유는 일반적인 다중회귀모형의 예측치를 S자 형태의 Bass 모형으로 보정하기 위함이다. 하지만 Bass 모형은 과거의 데이터를 기반으로 미래의 수요를 예측할 수 있는 모형이기 때문에, 미개봉 상태의 영화에는 적용할 수 없다. 미개봉 상태의 영화는 다중회귀모형을 보완하여 예측력을 높이는 방법을 연구할 필요가 있다. 가령 댓글 수와 평점평균만이 아닌 비정형 데이터인 댓글 자체 또한 변수로 추가할 수 있을 것이다. 추후에는 과거 데이터가 존재하지 않는 영화의 예측력을 높이기 위해 다양한 독립변수를 추가하여 분석을 진행해보는 것도 좋을 것이라 생각된다.

감사의 글

이 논문에 도움을 주신 박종민, 안영빈, 이윤선, 임보라, 조정아에게 감사의 말을 전하고자 합니다.

References

Bass, F. M. (1969). A new product growth for model consumer durables, *Management Science*, **15**, 215–227.

Jeon, S. and Son, Y. (2016). Prediction of box office using data mining, *The Korean Journal of Applied Statistics*, **29**, 1257–1270.

Jung, C., Cho, E., Moon, M., and Jung, Y. (2016). Movie demand diffusion pattern analysis: Applying bass diffusion model based on buzz amount of twitter. In *Proceedings of the 2016 Fall Conference of the Korea Society of Management Information Systems*, 111–115.

Jung, H. and Yang, H. (2013). Predicting financial success of a movie using multiple regression analysis, *Korea Society of Computer Information*, **21**, 275–278.

Korean Film Council (2016). 2015 Korean film industry settlement, *Korean Film.*

Lee, K. and Jang, U. (2006). Prediction of the box-office record of a movie using Bayesian choice model. In *Proceedings of the 2006 Spring Conference of the Korean Institute of Industrial Engineers*, **18**, 1428–1433.

Park, J., Chung, Y., and Cho, Y. (2015). Using the hierarchical linear model to Forecast Movie Box-office Performance: the effect of Online Word of Mouth, *Asia Pacific Journal of Information systems*, **25**, 563–578.

R Core Team (2014). R: A language and environment for statistical computing, *R Foundation for Statistical Computing*, Vienna, Austria, from: https://www.R-project.org/

Song, J. and Han, S. (2013). Predicting gross box-office revenue for domestic films, *Communications for Statistical Applications and Methods*, **20**, 301–309.

Talukdar, D., Sudhir, K., and Ainslie, A. (2002). Investigating new product diffusion across products and countries, *Marketing Science*, **21**, 97–114.

Venkatraman, N., Loh, L., and Koh, J. (1994). The adoption of Corporate governance mechanisms: a test of competing diffusion models, *Management Science*, **40**, 496–507.

Zhang, G. P. (2003). Time series forecasting using a hybrid ARIMA and neural network model, *Neurocomputing*, **50**, 159–175.

GLS와 Bass 모형을 결합한 하이브리드 모형을 이용한 영화 관객 수 예측

김보경[a] · 임창원[a,1]

[a]중앙대학교 응용통계학과

(2018년 3월 5일 접수, 2018년 5월 24일 수정, 2018년 6월 20일 채택)

요 약

국내 영화 산업 매출은 매년 증가하고 있다. 극장은 영화의 1차 판매 경로이며, 극장을 이용하는 관객 수는 부가판권에 영향을 준다. 따라서 극장을 이용하는 관객의 수는 영화 산업 매출에 직결되는 중요한 요소이다. 본 논문에서 특정일의 관객 수를 예측하기 위하여 다중선형회귀모형과 Bass 모형을 결합한 Hybrid 모형을 고려한다. 두 모형을 결합함으로써 회귀분석의 예측값을 Bass 모형의 예측값으로 보정하였다. 분석에는 개봉일이 모두 다른 세 영화를 이용하였다. All subset regression 방법을 이용해 모든 가능한 조합을 생성하고 5중 교차검증(5-fold cross validation)을 통해 5번 모형을 추정한다. 이 때 제곱근평균오차가 가장 작은 모형으로 예측값을 구한 뒤 Bass 모형의 예측값과 결합해 최종 예측값을 구하게 된다. 과거데이터가 존재할수록 Bass 모형의 가중치는 증가하면서 예측값에 보정효과를 준다는 것을 확인할 수 있었다.

주요용0어: 영화진흥위원회, 일반화최소제곱법, 다중회귀모형, Bass 모형, Hybrid 모형

이 논문은 2017년도 정부(과학기술정보통신부)의 재원으로 한국연구재단-차세대정보·컴퓨팅기술개발사업의 지원을 받아 수행된 연구임 (No. 2017M3C4A7083281).

[1]교신저자: (06974) 서울특별시 동작구 흑석로 84, 중앙대학교 응용통계학과. E-mail: clim@cau.ac.kr