The Korean Journal of Applied Statistics (2013) DOI: http://dx.doi.org/10.5351/KJAS.2013.26.3.441
26(3), 441–452

A Study for the Drivers of Movie Box-office Performance

Yon Hyong Kim[a,1] · Jeong Han Hong[b]

[a]Department of Statistics, Jeonju University; [b]Taylor Nelson Sofres Korea

(Received March 18, 2013; Revised June 4, 2013; Accepted June 4, 2013)

Abstract

This study analyzed the relationship between key film and a box office record success factors based on movies released in the first quarter of 2013 in Korea. An over-fitting problem can happen if there are too many explanatory variables inserted to regression model; in addition, there is a risk that the estimator is instable when there is multi-collinearity among the explanatory variables. For this reason, optimal variable selection based on high explanatory variables in box-office performance is of importance. Among the numerous ways to select variables, LASSO estimation applied by a generalized linear model has the smallest prediction error that can efficiently and quickly find variables with the highest explanatory power to box-office performance in order.

Keywords: Box office, generalized linear model, shrinkage estimation, variable selection.

1. 서론

국내 영화산업에 대한 학문적 연구는 영화의 흥행결과에 영향을 미치는 요인들을 규명하는데 주목해 왔으며, 대부분의 영화흥행 연구는 기존의 선행연구에서 사용되었던 요인체계를 바탕으로 변수들이 선택되었다. 영화의 속성 이론에 관련된 개별 요인이 흥행에 미치는 영향력을 검증하는데 집중함으로써, 상업적 시각에서 다양한 요인들을 통합적으로 이해한 영화의 흥행성과 모형을 개발하는 연구는 그리 활발히 진행되고 있지 않은 실정이다.

회귀분석모형에서 투입되는 설명변수가 많을 경우 과대적합(over-fitting) 문제가 발생할 수 있고, 더구나 설명변수 간 다중공선성이 있을 때에는 추정량이 불안정하게 될 위험이 있다. 모형의 복잡도와 예측력의 문제를 동시에 고려해야 하는 상황에서 흥행성과에 설명력이 높은 변수 위주로 최적의 변수 선택을 하는 것이 중요하나, 기존의 연구들은 관심변수의 유의성과 설명력에 주로 초점을 맞춤으로써 모형의 예측력과 타당성의 확보란 측면에서 간명하고 적합한 모형이 제시되기 어려웠다.

이 연구에서는 2013년 1월부터 3월까지 국내에서 개봉된 상업영화를 대상으로 영화흥행성과에 핵심적인 영향 요인을 회귀모형을 통해 살펴보았다.

회귀모형의 계수추정 방법에는 여러 가지가 있으나, 이 연구에서는 일반적인 Stepwise 회귀 외에 모형의 복잡도에 벌점을 주는 기법으로서 Ridge회귀 및 LASSO회귀를 이용하였다.

[1]Corresponding author: Professor, Department of Statistics, Jeonju University, Jeonju-Si, Jeollabuk-Do 560-759, Korea. E-mail: yhkim@jj.ac.kr

본 연구의 목적은 첫째, 기존의 영화흥행요인들로 밝혀진 변수들을 통합적으로 분석하여 예측오차가 가장 적고 흥행성과에 설명력이 가장 높은 변수 순으로 의미 있는 독립변수들을 빠르고 효율적으로 선택하는 것이다. 둘째, 상업적 시각에서 영화 개봉 전과 영화 개봉 후 온라인 구전의 효과를 실증 분석하는 것이다. 셋째, 연구에서 다루지 못했던 상영포맷을 고려하여, 영화 흥행 영향 요인으로서의 중요도를 검증하고자 한다.

2. 연구방법

2.1. 영화흥행성과의 연구모형

Eliashberg 등 (2006)은 영화산업의 가치사슬(value chain)에 있어 핵심단계를 제작(production), 배급(distribution), 상영(exhibition)의 3단계로 구분하였다.

상업적 시각에서 제작, 배급, 상영 각 단계별로 영화흥행예측과 관련된 연구문제를 다음과 같이 정리하여 볼 수 있다.

제작 단계: 영화흥행예측모형이 초기지표-대본, 캐스팅, 예상상영등급-를 토대로 얼마나 정확하게 개발될 수 있을 것인가?

배급 단계: 영화마케팅을 위한 예산을 다양한 미디어-전통적 미디어, 온라인 미디어-에 어떻게 배분하는게 최적인가? 영화 흥행에 온라인 리뷰, 온라인 구전이 얼마나 영향을 미치는가?

상영 단계: 최적의 스크린 수를 결정하기 위해 어떤 요인을 고려해야 하는가?

한국은 2000년대 이후 멀티플렉스 확산으로 전국 동시 개봉이 일반화되어 마케팅비용과 개봉 스크린 수가 관객동원에 영향을 미치는 요소로 간주되어 왔다. 인터넷이 중요한 커뮤니케이션 매체로 등장하면서 온라인 구전은 어떤 미디어보다도 강력한 구전커뮤니케이션 채널로서 소비자에게 위험을 회피할 수 있는 정보의 중요한 원천으로 작용하고 있다. 최근 들어 급속히 유행하는 3D, 4D 영화의 경우, 영화제작자, 배급사 그리고 대형 멀티플렉스영화관 입장에서 볼 때 동일한 관객으로부터 고객의 욕구를 자극해 더 높은 영화관람료를 지출하게 만드는 up-selling 동기요소로 작용하고 있다.

본 연구는 영화흥행의 영향요인을 규명하고자 다음과 같이 연구 문제를 설정하였다.

연구 문제 1: 개봉 전·후 온라인 평점과 빈도는 영화 흥행성과에 영향이 있는가?

연구 문제 2: 3D·4D 상영포맷은 영화 흥행성과에 영향을 미치는가?

2.2. 흥행성과의 정의

상업적 측면에서 영화의 흥행성과는 투자수익(ROI)의 관점에서 보면 제작비와 제반 경비를 초과하여 회수한 전국 매출액이라고 할 수 있다. 전국 매출액은 영화진흥위원회의 영화관입장권통합전산망을 통해 접근이 가능하지만 개별 영화의 제작비는 정확한 정보를 수집하기가 현실적으로는 불가능하다.

또한, 통합전산망의 전국 매출액은 제작자가 배급사와 극장 측에게 지불해야 할 배급수수료와 상영료의 배분이 되기 전에 관객들이 지불한 티켓 가격을 순수하게 합산한 것이다. 영화 산업의 주체 중 제작사 입장에서는 배급사, 극장 측과 배급수수료, 상영료를 배분하고 나서야, 제작사의 최종 매출이 되는 것이지만, 제작사, 배급사, 극장 상호간의 정확한 배분금액을 알아내는 것은 현실적으로 불가능에 가깝다

따라서 기존의 선행 연구들에 있어서는 전국 관객 수를 해당 영화의 흥행성과로 정의하였는데, 이는 소비자가 영화에 대해 지불하는 티켓 가격이 영화 별로 동일한 극장 산업 특성상, 관객 수와 매출액은 비

Table 2.1. Definition of variables

구분			변수	변수 형태	제작전 투자	개봉전 배급	개봉후 상영
독립변수	영화 속성	국적	한국, 미국, 기타	더미	o	o	o
		장르	코미디, 액션, 스릴러, 멜로, 드라마, 공포, 기타	더미	o	o	o
		관람 등급	전체관람가, 12세이상, 15세이상, 청소년관람불가	순서형	o	o	o
		감독 효과	감독작품 개봉이전 3년 감독영화의 매출액평균	척도형	o	o	o
		배우 효과	주연작품 개봉이전 3년 주연영화의 매출액평균	척도형	o	o	o
	경쟁 요소	배급사파워	배급사작품 개봉이전 3년 배급영화의 매출액평균	척도형		o	o
		스크린 수	전국 개봉 스크린 수	척도형		o	o
		상영 포맷	2D(일반)를 제외하고 3D, 4D, IMAX의 매출액 점유비중	척도형		o	o
	구전 효과	온라인평점	네이버의 일반인 평가의 평균 평점 (10점 만점)	척도형			o
		온라인빈도	네이버의 일반인 평가의 빈도	척도형			o
종속변수		매출액	영화진흥위원회 영화상영관 입장권 통합전산망집계 전국 매출액	척도형			

례한다고 할 수 있으므로 관객 수를 흥행성과로 정의한다고 해도 큰 무리가 없었다고 볼 수 있다. 그러나 최근 들어 급속히 유행하는 3D, 4D 영화의 경우, 2D(일반)보다 티켓가격이 비싸 객단가가 높기 때문에 관객 수 보다 매출액으로 흥행성과를 정의하는 것이 더 타당하다고 할 수 있다.

2.3. 흥행성과의 영향변수

영화 흥행에 관한 변수들은 크게 영화의 내적 요인과 외적 요인으로 구분되고 있다. 영화의 외적 요인은 다시 구전커뮤니케이션 영역과 배급유통경쟁영역으로 나누어 진다.

영화 흥행에 대한 영화의 내적 요인의 영향을 분석하기 위해 해당 영화의 국적과 장르는 더미변수를, 관람등급은 순서형 변수를 이용하였다. 또한 영화 흥행에 기여하는 감독과 배우의 효과를 측정하기 위하여 감독이 해당 작품이 개봉되기 직전 3년 동안 감독 또는 주연한 영화들의 매출액 평균을 이용하였다.

영화 흥행의 외적 요인 중 구전커뮤니케이션 영역은 온라인 평점과 빈도를 이용하였다. 배급유통경쟁영역 가운데 배급사 파워는 해당 영화의 배급사가 개봉시점으로부터 3년 전까지 배급한 영화가 동원한 매출액 평균을, 상영포맷은 3D, 4D, IMAX상영이 매출액에서 차지하는 비중(%)을 이용하였다.

이 연구에서는 한국의 영화시장에서 흥행성과에 영향을 미치는 요인들을 선행연구를 참고하여 국적, 영화 장르, 관람등급, 감독과 배우의 스타 파워, 배급사 파워, 스크린 수, 온라인 평점과 평가 빈도로 정하였다. 여기에 선행 연구에서는 다루지 않았던 영화상영포맷을 이 연구에 추가하여 Table 2.1에서 보는 바와 같이 총 10개의 설명변수를 고려하였다.

영화의 흥행을 나타내는 변수로는 영화진흥위원회 Box Office 전국매출액 통계를 이용하였으며, Table 2.1에 각각의 변수의 정의를 설명하였다. 제작 전 투자결정 단계에서 고려할 수 있는 요인과 개봉 전 최적편성 단계에서 고려할 수 있는 요인들을 O로 분류하였다.

영화제작 단계에서 배급사나 창투사가 투자자의 형태로 영화제작에 관여하게 되는데, 투자단계에서 고려할 수 있는 요인은 국적, 대본의 완결성, 장르와 예상관람등급, 감독과 배우의 캐스팅 등 영화의 내적 요인이다. 또한, 완성된 영화의 배급상영 단계에서 극장체인은 영화의 내적요인 이외에 구전커뮤니케이션, 배급사의 마케팅능력, 개봉 스크린 수 등과 같은 영화외적요인의 경쟁력을 함께 고려하여 최적편성을 시도하게 된다.

Table 3.1. Descriptive statistics of variables

변수	표본수	최소	최대	평균	표준편차
매출액	54	-	86,202,006,670	6,059,674,571	14,435,229,813
감독 효과	54	-	23,428,605,500	1,614,962,454	4,910,158,370
배우 효과	54	-	2,722,403	118,400	434,040.3
배급사 파워	54	-	10,276,962,523	4,301,966,077	3,890,856,371
스크린 수	54	13	894	275	219.5553
상영 포맷	54	0	66.5	4.717	11.1673
온라인평점(개봉전)	54	6.72	9.87	8.72	0.6924334
온라인빈도(개봉전)	54	12	1,444	380.5	379.8802
온라인평점(개봉후)	54	6.17	9.75	8.163	0.8324417
온라인빈도(개봉후)	54	14	14530	1,420	2626.023

Table 3.2. Revenue distribution of movies released in the first quarter of 2013 in Korea

그룹	기준	영화수	퍼센트	누적퍼센트
1	30억 원 미만	38	70.4	70.4
2	30억 원 이상 ~ 80억 원 미만	8	14.8	85.2
3	80억 원 이상 ~ 200억 원 미만	3	5.6	90.7
4	200억 원 이상 ~ 400억 원 미만	3	5.6	96.3
5	400억 원 이상	2	3.7	100.0
Total		54	100.0	

영화흥행성과에 영향을 미치는 요인을 규명하고자 영화속성, 경쟁요소, 구전효과를 고려하여 아래와 같은 모형으로 제시하였다.

$$\begin{aligned} \text{매출액} = {} & \text{국적} + \text{장르} + \text{등급} + \text{감독 효과} + \text{배우 효과} + \text{배급사 파워} + \text{스크린수} \\ & + \text{상영 포맷} + \text{온라인평점} + \text{온라인빈도}. \end{aligned}$$

3. 실증분석

3.1. 기술 통계

본 연구에서는 2013년 1월부터 2013년 3월까지 한국에서 개봉된 상업영화 54편을 실증분석에 이용하였다. Table 3.1은 본 연구에서 사용한 주요 변수에 대한 기술통계량이며, 분석대상에 이용된 54편의 평균 매출액은 60.5억 원으로 나타났다.

본 절에서는 회귀분석을 수행하기에 앞서, 설명변수와 종속변수들에 대한 탐색적 기초통계분석 및 가공을 수행하였다. 통계적 모형은 자료에 대한 정규분포 가정에 기반한 추론이 이루어지므로 변수 값의 분포가 정상분포의 형태에서 벗어나는 변수들에 대해서는 로그변환을 시도하였다.

Table 3.2를 보면 영화 매출액의 편차가 매우 크다는 것을 알 수 있는데, 200억 원 이상의 매출액을 기록한 흥행영화는 전체 개봉영화 편수의 9.3%에 불과할 정도로 그 숫자가 매우 적은 것을 볼 수 있다.

3.2. 상관성의 검토

Table 3.3은 더미변수를 제외한 주요 변수 간, 즉 종속변수와 모형에 포함한 독립변수들과의 스피어만 상관분석을 실시한 결과이다. 매출액과 감독 효과, 배급사 파워, 스크린 수, 온라인평점(개봉 전), 온라

Table 3.3. Correlation between variables

Spearman의 상관계수		매출액	감독	배우	배급사	스크린수	상영포맷	온라인평점 개봉전	온라인빈도 개봉전	온라인평점 개봉후	온라인빈도 개봉후
매출액	상관계수	1.000	.311	.114	.643	.956	.397	.875	.641	.171	.875
	유의확률(양측)	·	.022	.411	.000	.000	.003	.000	.000	.217	.000
	N	54	54	54	54	54	54	54	54	54	54
감독	상관계수	.311	1.000	.697	.390	.339	−.161	.404	.429	.103	.404
	유의확률(양측)	0.22	·	.000	.004	.012	.245	.002	.001	.457	.002
	N	54	54	54	54	54	54	54	54	54	54
배우	상관계수	.114	.697	1.000	.302	.164	−.160	.181	.229	.113	.181
	유의확률(양측)	.411	.000	·	.026	.236	.248	.189	.096	.416	.189
	N	54	54	54	54	54	54	54	54	54	54
배급사	상관계수	.643	.390	.302	1.000	.643	.213	.618	.571	.143	.618
	유의확률(양측)	.000	.004	.026	·	.000	.122	.000	.000	.303	.000
	N	54	54	54	54	54	54	54	54	54	54
스크린수	상관계수	.956	.339	.164	.643	1.000	.390	.873	.624	.099	.873
	유의확률(양측)	.000	.012	.236	.000	·	.004	.000	.000	.475	.000
	N	54	54	54	54	54	54	54	54	54	54
상영포맷	상관계수	.397	−.161	−.160	.213	.390	1.000	.175	.021	.016	.175
	유의확률(양측)	.003	.245	.248	.122	.004	·	.207	.878	.909	.207
	N	54	54	54	54	54	54	54	54	54	54
온라인평점 개봉전	상관계수	.875	.404	.181	.618	.873	.175	1.000	.797	.176	1.000
	유의확률(양측)	.000	.002	.189	.000	.000	.207	·	.000	.204	.
	N	54	54	54	54	54	54	54	54	54	54
온라인빈도 개봉전	상관계수	.641	.429	.229	.571	.624	.021	.797	1.000	.260	.797
	유의확률(양측)	.000	.001	.096	.000	.000	.878	.000	·	.058	.000
	N	54	54	54	54	54	54	54	54	54	54
온라인평점 개봉후	상관계수	.171	.103	.113	.143	.099	.016	.176	.260	1.000	.176
	유의확률(양측)	.217	.457	.416	.303	.475	.909	.204	.058	·	.204
	N	54	54	54	54	54	54	54	54	54	54
온라인빈도 개봉후	상관계수	.875	.404	.181	.618	.873	.175	1.000	.797	.176	1.000
	유의확률(양측)	.000	.002	.189	.000	.000	.207	·	.000	.204	·
	N	54	54	54	54	54	54	54	54	54	54

인빈도는 유의미한 상관관계를 보인 반면, 매출액과 배우 효과, 온라인평점(개봉 후) 간에는 상관관계가 없는 것으로 나타났다.

주연배우가 해당 작품의 개봉 이전 3년 동안 주연한 영화의 평균 동원 관객 수를 나타내는 배우 효과는 감독 효과, 스크린 수와 상관관계가 있음을 알 수 있다. 스타배우의 캐스팅은 상대적으로 고 제작비 영화에서 이루어지는데, 제작비가 많이 든 블록버스터 영화일수록 파워가 큰 배급사를 통해 스크린 수를 많이 확보함으로써 투자금액을 조기 회수하려하기 때문으로 해석할 수 있다. 그러나 배우 효과는 온라인평점, 온라인빈도와는 유의한 상관관계를 보이지 않았다.

스크린 수는 감독 효과, 배급사 파워, 상영 포맷, 온라인평점(개봉 전), 온라인빈도와 상관성이 있으며 온라인빈도 는 감독 효과, 배급사 파워, 스크린 수와 상관관계가 나타났다. 온라인평점(개봉 전)은 온라인빈도(개봉 전)와 상관관계를 보이고 있는 반면, 온라인평점(개봉 후)과는 뚜렷한 상관관계를 보이지 않았다. 한편 온라인빈도(개봉 전)는 온라인빈도(개봉 후)와 상관관계를 나타냈다. 2차적으로 다중공선성을 판단하기 위해 VIF를 이용하였다. 변수들의 VIF가 1에서 10 미만의 값으로 나타나, 심각하지는 않지만 추가적으로 다중공선성을 의심해 볼 필요가 있는 것으로 나타났다.

Figure 3.1은 각 변수와 매출액 간의 산점도이다. 감독파워, 배우파워, 배급사파워의 그래프에서 X축 좌표 0의 값에 몰려있는 것을 발견할 수 있다. 이는 영화 개봉 이전 3년 기간 평균 동원관객수가 많은

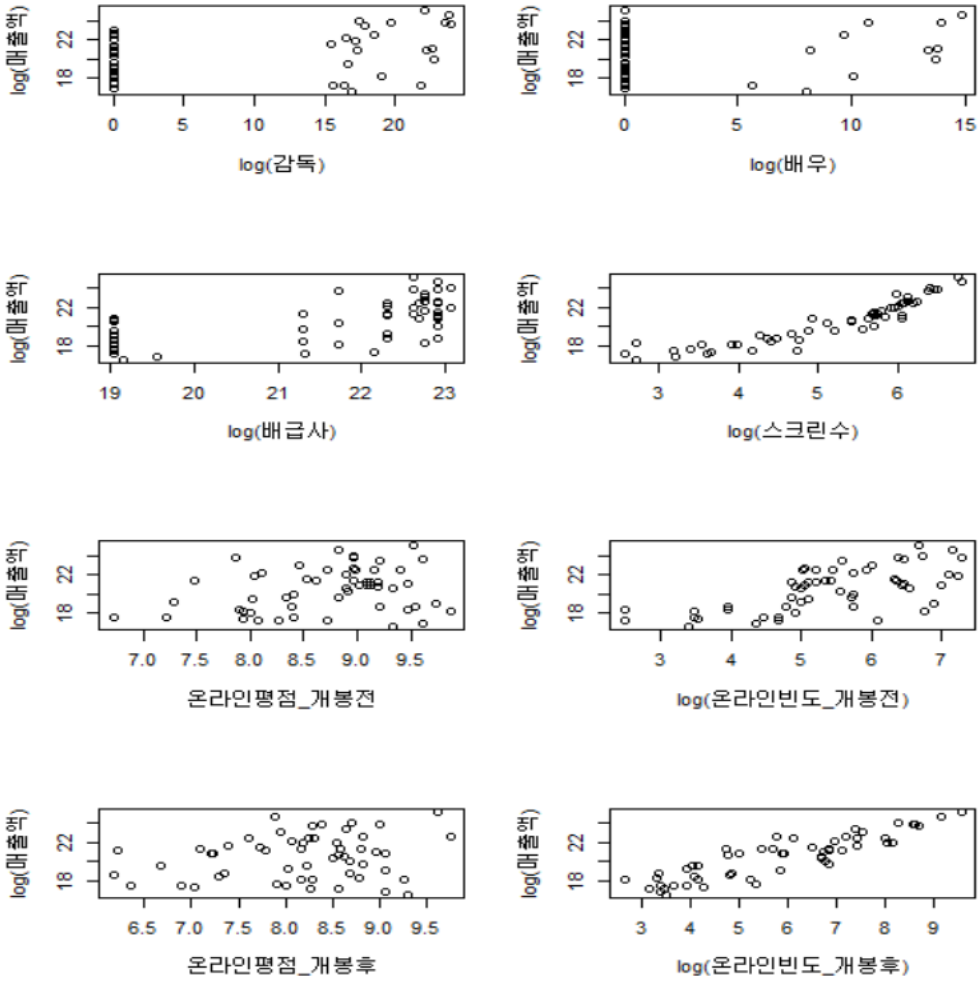


Figure 3.1. Scatter plots of variables

흥행작이 있는 감독과 배우라고 해서 항상 관객동원에 성공하는 것이 아니듯, 전체 개봉작 중 상당한 비중을 차지하는 무명 감독이나 신인 배우의 작품이라도 작품 여하에 따라 흥행에 성공할 수도 실패할 수도 있음을 보여주는 것이다. 또한, 이전 3년 기간 평균 동원관객수가 많은 배급사라고 해서 반드시 흥행에 성공하는 것이 아니고, 군소회사가 배급하더라도 작품 여하에 따라 관객동원에 성공할 수도 실패할 수도 있음을 보여준다.

3.3. 모형 비교

연구자는 연구방법의 편의를 위해 영화흥행에 관한 이론에 의해 증명되거나 가설로 검증할 필요가 있는 모든 변수를 모형에 포함시키는 변수선택방법(Enter method)을 우선 고려하게 된다. 그러나 현재 흥행성과에 영향을 주는 요인으로 모형에 포함되어 있는 변수라 하더라도, 통계적으로는 의미가 없는

변수가 있을 수 있다. 최소제곱법(OLS)을 이용하여 다중선형회귀분석을 실시하여 모수를 추정하는 경우, 설명변수의 개수가 증가하여 다중공선성(Multi-colinearity)이 존재하면 회귀계수의 분산이 커져서 회귀식의 예측력이 떨어지는 문제가 발생하게 된다. 따라서 적합한 변수를 선택하기 위해 통계적인 분석과정을 거쳐서 변수를 선별하여야 한다. 기존의 선행연구에서는 변수를 선택하기 위해 단계적 선택법(Stepwise selection)을 주로 활용해 왔다. 단계적 선택법으로 변수선택을 하더라도 선택된 설명변수간에 다중공선성이 있어 모수추정량의 분산이 팽창한다면 불안정한 추정이 되기 때문에 설명변수의 부분선택이 의미가 없게 된다. 예측모형에 있어서 종속변수에 영향을 주는 설명변수의 부분선택은 예측력이 높은 모형을 만드는데 중요한 역할을 한다. 최소제곱법(OLS)에서 출발하되, 특정 조건에서 어긋나는 경우 penalty를 주는 방식으로 모수를 안정적으로 추정하는 방법이 제안되었다.

Hoerl과 Kennard (1970)는 회귀계수 β에 대한 추정방법으로 최소제곱추정법 대신 아래와 같은 벌점화 기법을 제안하였다.

Ridge기법 추정치는 제약조건 $\sum_{j=0}^{p}\beta_j^2 \le t^2$하에서

$$\beta^{Ridge} = \arg\min_{\beta}\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}x_{ij}\beta_j\right)^2$$

로 주어진다. 라그랑즈 승수법(Lagrange Multiplier)에 의해

$$\beta^{Ridge} = \arg\min_{\beta}\left\{\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}x_{ij}\beta_j\right)^2 + \lambda\sum_{j=1}^{p}\beta_j^2\right\}, \quad t \ge 0,\ \lambda \ge 0$$

이 된다.

만약 $t = 0$이면 모형은 상수항만을 포함하고 $t = \infty$이면 최소제곱법과 동일하다.

그런데, Ridge기법이 축소 추정치를 주지만 회귀계수를 완전히 0으로 추정하지는 못하므로 변수선택이 여전히 어렵고 해석이 쉽지 않다. 즉, 많은 설명변수들 중 어떤 변수가 중요한 역할을 하는 지에 대한 판단이 그리 용이하지 않다.

Tibshirani (1996)는 회귀계수 절대값의 합이 주어진 상수보다 작게 하는 조건하에서 잔차제곱합을 최소화하는 LASSO(least absolute shrinkage and selection operator)기법을 제안하였다. LASSO기법은 Ridge기법처럼 최소제곱추정의 축소추정치를 줌과 동시에 설명력이 없는 설명변수들의 계수는 0으로 추정함으로써 자동적인 변수선택이 가능해지고 모형의 해석이 용이하게 된다.

LASSO기법 추정치는 제약조건 $\sum_{j=0}^{p}|\beta_j| \le t$하에서

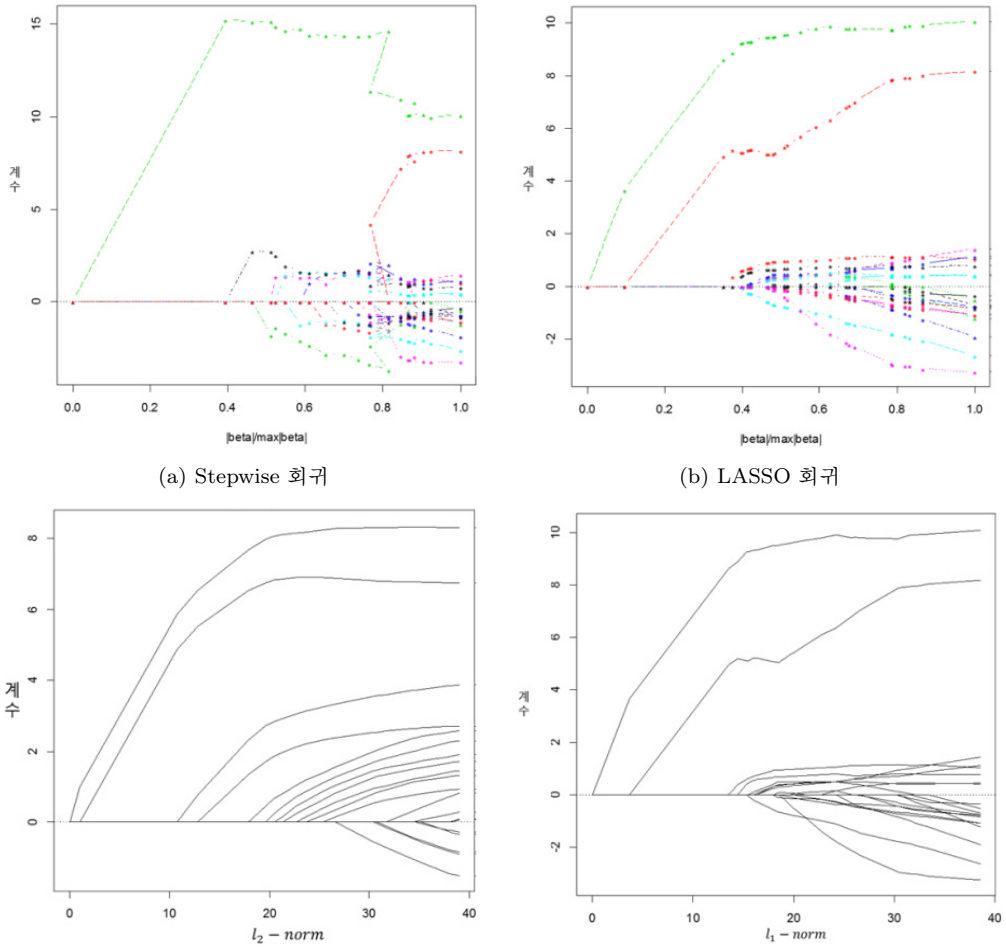
$$\beta^{LASSO} = \arg\min_{\beta}\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}x_{ij}\beta_j\right)^2$$

로 주어진다.

라그랑즈 승수법(Lagrange Multiplier)에 의해

$$\beta^{LASSO} = \arg\min_{\beta}\left\{\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}x_{ij}\beta_j\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j|\right\}, \quad t \ge 0,\ \lambda \ge 0$$

이 된다.



(a) Stepwise 회귀 (b) LASSO 회귀

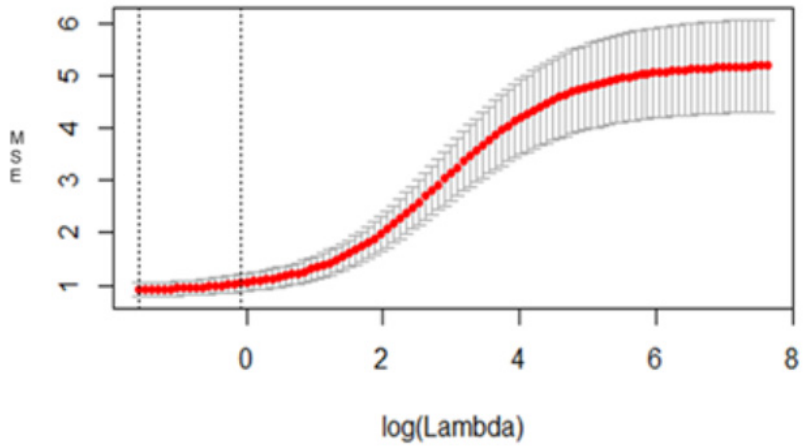
(c) Ridge 회귀 (d) LASSO 회귀

Figure 3.2. Model comparisons between Stepwise, Ridge and LASSO

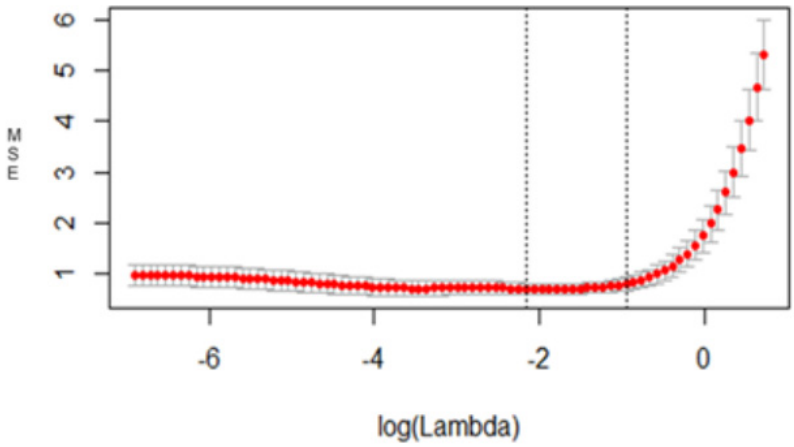
만약 $t = 0$이면 모형은 상수항만을 포함하고 $t = \infty$이면 최소제곱법과 동일하다. Ridge기법과 LASSO기법의 차이는 벌점이 $l_2 - \text{norm} \sum_{j=1}^{p} \beta_j^2$에서 $l_1 - \text{norm} \sum_{j=1}^{p} |\beta_j|$로 바뀐 점이다.

Figure 3.2의 (a)와 (b)는 각각 Stepwise와 LASSO 추정계수의 프로파일을 보여준다. 첫 번째 Stepwise회귀의 추정계수는 상당히 불안정한 반면 LASSO회귀의 추정계수는 안정적으로 나타나고 있다. Figure 3.2의 (c)와 (d)를 통해 Ridge와 LASSO 추정계수의 프로파일을 비교할 수 있다. Ridge회귀의 경우 모든 변수의 값이 0이 아니다. 반면 LASSO회귀의 경우 최적화된 벌점모수 값에 대하여 얻어진 변수 값들 중에서 일부는 0이다.

선형회귀분석의 가정이 충족되지 않는 문제점을 해결하는 방안의 하나인 일반화선형모형(GLM)에 벌점화기법을 적용할 수 있다. Ridge 및 LASSO제약조건 하의 최대우도(Penalized maximum likelihood)방법에 의해, 벌점모수(Penalty parameter)의 경로(path)와 교차확인법(Cross-validation)에 의해 선택된 모수 추정치를 구할 수 있다.



(a) Ridge 일반화 선형모형



(b) LASSO 일반화 선형모형

Figure 3.3. Trace plots of cross-validation errors (Ridge GLM and LASSO GLM)

Figure 3.3은 Ridge와 LASSO추정량의 조절모수로서 λ의 값의 변화에 따른 교차타당성오차(Cross-validation error)의 트레이스를 보여준다. LASSO추정의 경우 $\text{Log}(\lambda) = -2.100834$ 즉, $\lambda = 0.1223543$ 일 때 교차타당성오차가 최소가 됨을 알 수 있다.

Ridge기법은 축소추정치를 주지만 변수선택은 하지 않으므로 고차원 자료의 경우 최종 모형에 대한 해석이 용이하지 않다. 반면, LASSO기법은 축소추정과 변수선택을 통해 예측력을 향상시키는 동시에 최종 모형에 대한 해석을 용이하게 하는 방법이다. 변수의 개수가 증가하면 RSS는 감소하지만 Cp는 처음에는 감소하다 모형이 복잡해지면서 다시 증가하게 된다.

Table 3.4에서 보는 바와 같이 Cp 값이 14.489로 최소가 되는 9단계에서 모형을 선택할 수 있다. 선택한 모형에서 스크린 수 > 온라인빈도(개봉 후) > 배급사 파워 > 온라인평점(개봉 후) > 감독 효과 > 스릴러 > 상영 포맷 > 코미디 > 온라인평점(개봉 전)의 순으로 변수선택이 이루어졌다.

Table 3.4. RSS and Cp at each step of LASSO GLM

단계	Df	Rss	Cp	투입 변수
0	1	272.274	508.003	상수항
1	2	173.981	307.836	스크린 수
2	3	35.099	24.190	온라인빈도 (개봉 후)
3	4	31.388	18.558	배급사 파워
4	5	28.617	14.858	온라인평점 (개봉 후)
5	6	28.224	16.049	감독 효과
6	7	26.848	15.220	스릴러
7	8	26.403	16.304	상영 포맷
8	9	26.325	18.144	코미디
9	**10**	**23.575**	**14.489**	**온라인평점 (개봉 전)**
10	11	23.315	15.953	멜로
11	12	22.747	16.784	드라마
12	13	22.626	18.536	15세 이상 관람가
13	14	21.600	18.425	공포
14	15	21.272	19.752	온라인빈도 (개봉 전)
15	16	19.016	17.111	미국
16	17	18.215	17.465	배우 효과
17	18	17.352	17.689	액션
18	19	17.083	19.136	12세 이상 관람가
19	20	16.402	19.735	전체 관람가

온라인구전이 개봉 전보다는 개봉 후 각종 온라인매체를 통해 네티즌들의 입소문이 퍼지면서 영화흥행에 영향을 주는 요소임이 확인되었다. 또한 온라인 평가는 개봉 후평점과 빈도가 모두 흥행성과에 영향을 미치지만, 평점보다는 빈도가 훨씬 높은 영향력을 보이는 것으로 분석되었다. 반면, 개봉 전에는 온라인빈도는 유의하지 않고 온라인평점만 흥행성과와 유의미한 관계를 보인 것으로 나타났다. 온라인평점과 빈도를 모두 고려한 선행연구 (박승현 등, 2011; Kim과 Hong, 2011)에서 온라인빈도만 영향력이 유의한 것으로 나타난 것과 다소 다른 면이 있다.

Table 3.5는 Ridge와 LASSO회귀에 의한 회귀계수 추정결과를 비교한 것이다. LASSO회귀는 효과가 작아 의미가 없는 설명변수에 대한 회귀계수를 0으로 추정할 수 있어, 예측력(Prediction accuracy)이 높고 모형을 쉽게 해석할 수 있다는 점에서 매우 유용하고 robust함을 알 수 있다.

4. 결론

본 연구에서는 2013년 1월부터 2013년 3월까지 국내에서 상영된 상업영화를 대상으로 영화 흥행 결정요인을 파악하였다. 선행 연구 결과와 비교하기 위해 온라인구전의 영향력을 분석하였으며, 기존 연구에서 고려하지 못한 상영포맷의 영화흥행성과에 대한 영향력을 검증하였다.

상업적 시각에서는 다양한 유형의 많은 변수가 존재하기 때문에 회귀분석모형에 투입되는 설명변수가 많을 경우 과대적합(over-fitting) 문제가 발생할 수 있고, 설명변수 간 다중공선성이 있을 때에는 추정량이 불안정하게 될 위험이 있다.

이 연구에서는 LASSO회귀를 적용하여 스크린 수, 온라인빈도(개봉 후), 배급사 파워, 온라인평점(개봉 후), 감독 효과, 스릴러, 상영포맷, 코미디, 온라인평점(개봉 전)의 순으로 9개 변수를 흥행성과의 영향변수로 선택하였다.

Table 3.5. Comparisons of parameter estimates (Ridge GLM and LASSO GLM)

변수	Ridge 계수	LASSO 계수
상수항	8.573164525	9.818450466
한국	0.160606365	
미국	−0.112759978	
코미디	0.307067312	0.112886963
액션	−0.121811865	
스릴러	−0.575660286	−0.170964102
멜로	−0.565155617	
드라마	−0.104107346	
공포	−0.622058137	
전체	−0.028500248	
12세 이상 관람가	−0.040066967	
15세 이상 관람가	0.191984507	
감독 효과	0.011724245	0.002327929
배우 효과	−0.009289321	
배급사 파워	0.137600487	0.071290160
스크린 수	1.000623575	1.129656656
상영 포맷	0.013814369	0.003471540
온라인평점(개봉 전)	0.120868525	0.025617524
온라인빈도(개봉 전)	−0.143661762	
온라인평점(개봉 후)	0.103710157	0.102039446
온라인빈도(개봉 후)	0.459592455	0.385586314

영화 속성으로 장르-코미디, 스릴러, 감독 효과가 영화흥행성과에 유의한 것으로 나타났다. 구전효과로서 개봉 전에는 온라인평점이, 개봉 후에는 온라인 평점과 빈도 모두 영화 관객을 유인하는 요인으로 분석되었다. 경쟁 요소로는 스크린 수, 배급사 파워, 상영 포맷이 유의한 영향을 나타내는 것으로 나타났다. 상영 포맷 즉, 3D·4D는 본격적으로 디지털화되고 있는 영화제작 및 상영관 리노베이션 추세를 감안할 때 흥행성과인 매출액에 미치는 영향력이 점차 증대될 것으로 전망된다.

국내 영화 산업에서 상업적인 영화흥행예측은 스코어카드에 의한 평가점수합계로 이루어지고 있다. 스코어카드를 구성하는 항목들의 가중치가 주관적 경험에 근거한 배분으로 이루어지고 있는데, LASSO회귀를 통해 도출한 흥행결정요인들의 상대적 영향력을 가중치로 하여 각 요인에 대한 상영 예정 영화의 평가점수를 가중 평균한다면 보다 합리적이고 객관적으로 흥행성과를 예측할 수 있으리라 기대된다.

References

Eliashberg, J., Elberse, A. and Leenders, M. A. A. M. (2006). The Motion Picture Industry: Critical Issues in Practice, Current Research, and New Research Directions, *Marketing Science*, **25**, 638–661.

Hoerl, A. E. and Kenard, R. W. (1970). Ridge regression: Biased estimation for non-orthogonal problems, *Technometrics*, **2**, 55–68.

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**, 859–869.

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, *Journal of Communications Research*, **11**, 231–258.

Tibshirani, R. (1996). Regression and Shrinkage via lasso, *Journal of the Royal Statistical Society*, **58**, 267–288.

영화흥행 영향요인 선택에 관한 연구

김연형[a,1] · 홍정한[b]

[a]전주대학교 통계학과, [b]테일러넬슨 소프레스 코리아

(2013년 3월 18일 접수, 2013년 6월 4일 수정, 2013년 6월 4일 채택)

요 약

국내 영화 산업은 투자·배급사·멀티플렉스로 수직 계열화된 대기업 중심으로 온라인 구전 마케팅이 활발히 진행되고 있다, 최근에는 대기업 계열의 멀티플렉스 영화관 중심으로 3D·4D 영화포맷 복합상영을 통해 up-selling을 통한 흥행성과 극대화를 도모하고 있다. 영화산업 기술진보와 흥행여건 변화에 따라, 기존 관객 수 대신 매출액을 흥행성과로 정의하고, 국내 개봉 상업영화를 대상으로 축소추정기법을 포함한 여러 회귀모형을 적용하였다. 특히 LASSO회귀의 경우, 교차타당성 방법을 이용한 예측오차가 가장 적고 흥행성과에 설명력이 높은 변수 순으로 의미 있는 독립변수들을 빠르고 효율적으로 선택할 수 있었다. 2013년도 1분기 개봉 영화를 대상으로 실증분석 결과, 개봉 후 온라인 평점과 빈도 모두 영향력이 높았으나, 개봉 전에는 온라인 평점만 효과적인 것으로 나타났다. 상영포맷 또한 흥행성과에 유의한 영향을 미치는 것으로 나타났다.

주요용어: 영화흥행, 일반화선형모형, 축소추정, 변수선택.

[1]교신저자: (560-759) 전북 전주시 완산구 효자동3가 1200, 전주대학교 통계학과, 교수. E-mail: yhkim@jj.ac.kr