외화의 국내 흥행 요인 분석

20130602 백찬규

1. Problem motivation

영화는 대표적인 대중 예술이다. 순수한 의미의 예술 영화는 흔치 않으며 대부분이 대중의 여가 생활을 위한 상품으로써 소비된다. 상품이 된 영화는 철저히 시장의 논리를 따른다. 시장의 논리에서 '좋은 영화'는 곧 수익을 많이 창출한 영화이다. 그렇기에 경제적인 관점에서 '좋은 영화'를 판별하는 기준은 작품성보다는 상품성이 될 수 밖에 없다. 그리고 다분히 주관적인 작품성이라는 기준에 비해 상품성은 수치화 가능한 특징을 지니며, 이는 계량적 분석의 대상이 된다.

영화라는 상품을 다루기 위해서는 영화 산업 전반에 대한 이해도 물론 중요하다. 하지만 투자자 혹은 배급자의 입장에서는 개별 영화의 흥행 요인을 도출하는 것이 더 중요하다 할 수 있다. 투자 결정과 제작 그리고 홍보 단계까지 모든 과정의 평가가 개별 영화의 흥행이라는 하나의 요소에 좌우되기 때문이다. 한 작품을 시장에 내놓기 위해서는 상당한 예산과 시간이 소요된다. 그렇기에 개별 작품의 흥행 가능성을 오판하여 투자가 실패한다면 적잖은 손해를 감수해야 할 것이다.

영화의 흥행 여부를 판단하는 요소는 역시 관객수이다. 우선 많은 관객수는 높은 매출과 직결된다. 투자자 입장에서는 제작비를 보전할 수 있는 최소한의 관객수라도 확보하여 손익분기점을 넘기는 것이 간절할 것이다. 게다가 관객수는 영화의 영향력을 가늠하는 수단이 된다. 현대 컨텐츠 산업에서 많은 사람이 해당 작품을 관람했다는 사실은 그 자체로 상품성을 갖는다. 인지도가 높은 작품은 2차적인 유통, 혹은 작품을 활용한 판촉활동 등에 강점이 있기 때문이다. 예를 들면, 최근의 영화는 극장 상영이 종료된 후 VOD 다시 보기 서비스가 제공된다. 또는 작품 속 인물을 활용하여 캐릭터 산업으로의 확장이 일어나는 경우도 흔하다. 이러한 2차적인 수익을 가져오는 원동력이 바로 해당 컨텐츠의 영향력, 기본적으로 관객 수라고 할 수 있다.

관객수의 정확한 예측이 가능하다면 투자자 혹은 배급자가 떠안는 리스크를 크게 줄일 수 있을 것이다. 영화 한편을 기획, 제작하여 배급, 홍보하는 데 드는 비용은, 점점 커지는 영화산업의 몸집만큼 점점 더 늘어난다. 이는 영화 산업 관계자에게 큰 부담이 아닐

수 없다. 심지어 막대한 예산을 들인 영화의 경우에는 영화 한편의 흥행 실패가 곧 제작, 배급사의 경영 위기로까지 이어지기도 한다. 때문에 개별 영화의 흥행 가능성을 사전에 분석하는 것은 경영학적 관점에서 필수적이다.

2. Literature review

앞선 연구를 살펴보면 국내외를 막론하고 개별 영화의 흥행을 예측하는 것이 매우 어려운 일임을 강조한다. 그럼에도 영화 흥행 실적을 예측하는 작업은 70년 대 이후 학계에서 본격적으로 진행되었다. 8, 90년 대에는 회귀 분석적 방법이 활발하게 이루어져 영화 흥행에 영향을 주는 유의미한 변인으로 제작비, 개봉 스크린 규모, 전문가 평점 등이 확인되었다. 또한 국내의 연구는 한국 배우의 스타 파워, 사회 풍자성, 애국심 유발 등 한국 정서에 알맞게 변인을 조정하여 분석 대상으로 삼았다. 그러나 대부분의 연구가 한정된 기간에 이루어지거나 단편적인 변수만을 고려하는 등, 완벽하게 객관적인 지표로 수행된 효과적인 연구는 행해지지 않았다는 평이다[1]

3. Statement of research objectives

해당 연구는 외화의 국내 관객수를 예측하는데 초점을 두었다. 이는 해외에서 먼저 개봉한 영화를 국내에 배급하는 경우를 종종 볼 수 있기 때문이다. 또한 수집한 데이터의 한계 때문이기도 하다. 여러 제약상 국내외 영화 관련 정보가 총 망라된 데이터를 찾을 수 없었으며, 수집한 데이터를 정제하고 활용하는 데에도 한계가 있었다. 게다가 분석 방법 역시 회귀분석으로 제한적이었기에 회귀분석에 적합한 데이터만이 활용되었다. 그럼에도 불구하고 적절한 국외의 데이터를 국내 데이터와 병합할 수 있었으며, 국내 흥행 요인에 대한 통찰을 제공하기에 충분했다고 판단하였다.

연구의 주된 목적은 회귀분석을 이용하여 관객 수라는 종속 변수를 예측하는 것이다. 데이터 분석의 목적은 결국 예측이 되어야 한다. 단순히 과거 사실을 알려주는 연구로는 언제나 뒷북만 울릴 뿐이다. 예측을 내놓아 실제 의사결정에 영향을 줄 수 있어야 분석이 실체적 의의를 지닌다. 그래서 독립 변수 선정 시 예측 불가한 사후적인 관측치는 제외하였다. 예를 들면 총 매출을 나타내는 변수는 관객 수와 밀접한 상관이 있지만 상영이 끝난 뒤에 얻게 되는 결과이므로 이용하지 않았다.

---

[1] 한국 영화 시장의 흥행 결정 요인에 관한 연구_박승현, 정완규(2009)

4. Description of data and applied methodology

우선 관객수를 결정하는 여러 요인을 상정했다. 기본적으로 생각할 수 있는 변수로는 상영되는 스크린의 수, 작품성, 영화에 투입된 예산, 출연 배우의 인지도 등이 있다. 또한 영화의 장르, 제작 국가 상영 등급과 같은 특성들도 있다. 그 외에도 상영 연도나 상영 계절 혹은 상영시간 등의 시간적 변수들도 영향을 줄 수 있을 것이다.

상기한 데이터를 포괄하는 원 자료를 얻기 위해 국내 데이터와 해외 데이터를 합쳐야 했다. 우선 한국영화 데이터베이스(Korean Movie Database, 약칭 KMDb)에서 개봉작 흥행 순위 정보를 찾았다. 그리고 외화 관객수와 전국 스크린 수를 분리해냈다. 다음으로 해외 개봉 시의 데이터를 인터넷 영화 데이터베이스(Internet Movie Database, 약칭 IMDb)에서 가져왔다. 자료가 웹 페이지 형식으로 표시되어 데이터를 추출하는 것이 쉽지 않았다. 다행히 'Kaggle'이라는 데이터베이스 플랫폼에서 'IMDB 5000 Movie Dataset'이라는 정형화된 오픈 데이터 셋을 제공하였기에 분석에 이용할 수 있었다.

원 자료는 5043행 28열의 방대한 자료였으나 국내 데이터와 병합 과정에서 대부분의 관측치가 제외되었다. 올바른 분석을 위해서는 예상된 모든 독립 변수에 대해 온전한 관측치가 존재해야 하지만 불행히도 그렇지 못했기 때문이다. 해외와 국내의 데이터 각각이 불완전한 경우가 많았기에, 병합된 데이터의 양은 줄어들 수 밖에 없었다. 예를 들면, 개봉일 기준 국내 스크린 수 집계가 가능해진 시기는 2008년 이후였기에 그 이전 데이터는 사용할 수 없었다. 또한 해외 데이터의 추출 시점이 2016년 8월이었기에 그 이후의 데이터도 제외하였다. 또한 가정된 독립 변수 중 결측치가 존재하는 데이터도 사용할 수 없었다. 게다가 분석의 대상이 상업영화이기에 어느 정도 상업성을 갖추었다고 인정되는 수준, 구체적으로 국내 관객수 100만명 이상의 영화만을 이용하였다. 그 결과 총 175행 11열의 데이터를 확보하였다.

종속변수로는 국내 관객수를 선정하였다. 한국영화 데이터베이스 기준으로 하였으며 전국 관객 합계를 찾을 수 없는 경우 배제하였다. 그리고 원 자료의 28개 독립변수 중 불필요하거나, 중복 가능성 있는 자료를 제외하였다. 예를 들면, 흑백영화를 구분하는 color변수는 08년 이후 데이터에서 무의미하기에 삭제하였고, 감독과 배우의 이름 등의 변수 역시 삭제하였다. 그리고 출연진의 facebook likes를 모두 합하여 작품의 인지도를 대표하는 변수로 설정하였다. 또한 예측에 중점을 둔 연구인 만큼 영화의 총 매출을 의미하는 gross 변수를 제외하는 등, 예측이 불가능한 사후적인 자료를 삭제하였다. 그 외에도 num_critic_for_review, num_voted_users 등 변수의 영향은 작품성과 인지도 변수에 충분히 반영되었다고 판단하여 배제하였다

$$관객수 = \beta1 + \beta2*스크린\ 수 + \beta3*상영연도 + \beta4*작품성 + \beta5*배우\ 인지도 + \beta6*예산 + \beta7*상영시즌 + \beta8*장르 + \beta9*상영등급 + \beta10*제작국가$$

위와 같은 선형회귀식을 가정하였다. 개별 변수를 짚어 보자면, 우선 스크린 수는 영화가 상영된 스크린의 개수를 의미하며 관객수와 밀접한 양의 상관관계가 있을 것이라 추측했다. 다음으로 상영 연도는 영화의 개봉 연도를 의미한다. 해가 갈수록 영화산업의 규모가 확장되고 있기에 역시 양의 상관관계를 보일 것으로 여겨졌다. 이어서 작품성과 배우 인지도는 관객수 예측의 핵심적인 변수가 될 것으로 예상했다. 수치화하기 어려운 부분이 있어 변수를 선택하는 것에 제약이 있었지만, 충분히 객관성을 확보한다면 상당히 유의미한 정보라고 생각했다. 상식적으로 양의 부호를 보이리라 여겨졌다. 다음으로 예산은 영화의 제작비를 의미한다. 역시 관객수와 양의 상관을 보일 것 같았다.

남은 네 개의 변수, 상영시즌, 장르, 상영등급, 제작국가는 categorical variable이다. 일반적으로 영화의 흥행을 좌우 할만한 categorical variable을 생각해보았다. 첫째, 상영시즌이 관객수에 영향을 준다고 판단했다. 상영시즌은 영화를 상영한 주된 계절을 의미한다. 영화는 실내에서 상영되기에 날씨의 영향을 크게 받지 않는다. 바꾸어 말하면 날씨가 너무 춥거나 더울 때를 가리지 않고 언제나 좋은 선택지가 될 수 있다. 따라서 여름과 겨울철에 날씨 영향을 크게 받는 실외 활동보다 인기가 높을 것으로 예측했다. 둘째, 영화의 장르가 관객수에 영향을 미칠 것으로 예상했다. 때로는 다른 요인과 관계없이 영화의 장르만으로 극장에서 관람 여부가 갈리곤 한다. 즉, '영화관에서 볼 만한 영화'와 '집에서 봐도 괜찮은 영화'의 구분이 존재한다. 이는 꽤나 보편적인 추측이라 생각했는데 대부분의 관객이 '극장용' 영화로 스케일이 크고, 볼거리가 많은 액션 영화를 선호하는 경향을 보이기 때문이다. 반면 액션 장르가 아닌 영화에 대해서는 상대적으로 '극장용' 선호도가 떨어진다고 판단하였다.

셋째, 제한 상영 판정을 받은 영화는 관객수가 감소할 것이라 가정했다. 성인만이 관람 가능한 영화는 가족 단위 관객 혹은 청소년 관객을 놓치게 된다. 때문에 그렇지 않은 영화보다 관객수의 손실이 있을 것이라 생각하였다. 넷째, 제작 국가(특히 미국)에 따라 관객수의 증감이 있을 것이라 판단했다. 국가마다 문화가 다르기에 영화의 색채도 달라진다. 할리우드 식 블록버스터 영화에 익숙해진 국내 관객의 성향을 고려할 때, 미국에서 제작된 영화와 그렇지 않은 영화 간에 차이가 존재할 것이라 생각했다.

이상의 예측이 대부분 들어맞는다면 전체 모형의 설명력이 유의미 할 것이다. 즉, 모든 회귀 계수 값이 0이라는 귀무가설을 기각할 것이다. 나아가 모든 개별 변수의 설명력이 유의미하다면, 개별 회귀 계수의 값이 0이라는 귀무가설이 기각될 것이다.

실제 사용한 변수는 아래와 같다.

-screens: 상영된 스크린의 수를 의미한다.

-year: 상영연도를 의미한다.

-imdb_score: 영화에 대한 IMDb 이용자들의 평점을 평균 낸 수치이다. 작품성을 나타내는 변수로 사용 가능하다 판단하였다.

-facebook_likes: 출연진의 Facebook likes를 합산한 수치이다. 배우 인지도를 측정하는 변수로 적합하다 판단하였다. 다만 이 수치는 배우가 소개된 IMDb 페이지에 대한 likes이며 배우의 Facebook 계정과는 직접적 관련이 없다.

-budget: 작품의 제작비를 의미한다.

-season: binary variable로 상영 시즌을 의미한다. 1이면 여름/겨울, 0이면 봄/가을을 나타낸다.

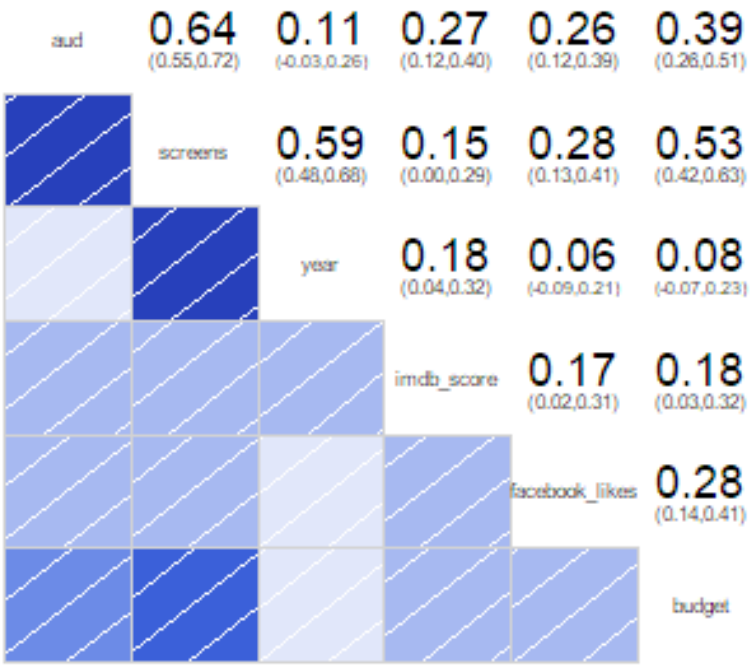
-genres: binary variable로 장르를 의미한다. 1이면 액션 영화, 0이면 그렇지 않은 영화를 나타낸다.

-rating: binary variable로 상영 등급을 의미한다. 1이면 제한상영등급(R or NC-17), 0이면 그렇지 않음을 나타낸다.

-country: binary variable로 제작 국가를 의미한다. 1이면 미국에서 제작된 영화, 0이면 그렇지 않은 영화를 나타낸다.

```
   title               aud              screens           year        imdb_score
Length:175         Min.   : 1006833   Min.   : 201.0   Min.   :2008   Min.   :4.200
Class :character   1st Qu.: 1461848   1st Qu.: 460.5   1st Qu.:2010   1st Qu.:6.350
Mode  :character   Median : 2182227   Median : 602.0   Median :2012   Median :7.000
                   Mean   : 2792156   Mean   : 683.4   Mean   :2012   Mean   :6.957
                   3rd Qu.: 3233911   3rd Qu.: 826.0   3rd Qu.:2014   3rd Qu.:7.700
                   Max.   :13624328   Max.   :1991.0   Max.   :2016   Max.   :9.000
facebook_likes       budget             season           genres        rating
Min.   :    13   Min.   :  3300000   Min.   :0.0000   Min.   :0.0   Min.   :0.0000
1st Qu.:  5026   1st Qu.: 75000000   1st Qu.:0.0000   1st Qu.:0.0   1st Qu.:0.0000
Median : 16967   Median :130000000   Median :1.0000   Median :1.0   Median :0.0000
Mean   : 22148   Mean   :128264000   Mean   :0.5543   Mean   :0.6   Mean   :0.1314
3rd Qu.: 29545   3rd Qu.:175000000   3rd Qu.:1.0000   3rd Qu.:1.0   3rd Qu.:0.0000
Max.   :106759   Max.   :260000000   Max.   :1.0000   Max.   :1.0   Max.   :1.0000
   country
Min.   :0.0000
1st Qu.:1.0000
Median :1.0000
Mean   :0.8514
3rd Qu.:1.0000
Max.   :1.0000
```

변수에 대한 기초통계량은 위와 같다.



다음으로 종속변수와 독립변수 간 상관분석을 진행하였다. 실질 변수의 상관 분석 결과 aud와 screens의 관계가 두드러진다(0.64). 때문에 회귀분석 시 주된 독립변수는 screens가 될 것이라 예상 가능하다. 하지만 Screens의 영향력은 독립변수 간 다중공선성의 문제를 낳는다. Year과 budget 간 상관계수가 각각 0.59, 0.53으로 높은 편이기 때문이다. 독립변수간 독립을 가정하는 회귀분석에서 이러한 다중공선성은 모형을 왜곡할 수 있는 위험이다.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2514187     231976  10.838   <2e-16 ***
season        501490     311584   1.609    0.109
---
```

Binary variable의 영향력을 가늠하기 위한 simple regression의 결과이다. season 변수가 1인 경우 97개, 0인 경우 78개가 관측되었고 영향력은 위와 같았다. 즉, 다른 변수가 통제되지 않았을 경우 여름/겨울철 관객수는 그렇지 않은 경우보다 약 501,490명 증가했으며 p-value 0.109 정도의 유의성을 보인다.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2163772     238865   9.059 2.76e-16 ***
genres       1047307     308373   3.396 0.000848 ***
---
```

genres 변수가 1인 경우 105개, 0인 경우 70개가 관측되었고 영향력은 위와 같았다. 즉, 다른 변수가 통제되지 않았을 경우 액션 영화 관객수는 비액션 영화보다 약 1,047,307명 증가했으며 p-value 0.0008 로 유의한 결과이다.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2923390     165138  17.703   <2e-16 ***
rating       -998522     455513  -2.192   0.0297 *
---
```

rating 변수가 1인 경우 23개, 0인 경우 152개가 관측되었고 영향력은 위와 같았다. 즉, 다른 변수가 통제되지 않았을 경우 제한 상영 등급 영화의 관객수는 그렇지 않은 경우보다 약 998,522명 감소했으며 p-value 0.0297로 유의한 결과이다.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2331818     403009   5.786 3.31e-08 ***
country       540666     436758   1.238    0.217
---
```

country 변수가 1인 경우 149개, 0인 경우 26개가 관측되었고 영향력은 위와 같았다. 즉, 다른 변수가 통제되지 않았을 경우 미국 영화의 관객수는 비 미국영화보다 약 540,666명 증가했으며 p-value 0.217로 유의하지 않다. 관측치가 구분되는 경우도 적고, 결과도 유의하지 않아 분석에서 country 변수를 제외하기로 결정했다.

처음 가정된 회귀 식은 다음과 같았다.

Aud = screens year imdb_score facebook_likes budget season genres rating

그러나 위 모형은 변수 간 다중공선성이 존재하는 불완전한 모형이다. 따라서 모형을 수정해야 한다. 변수 중 screens와 year, screens와 budget 간 교호작용이 존재한다. 이러한 경우 다중공선성을 야기하는 screens 변수를 모형에서 제외하거나 변환하여 사용하는 것이 일반적이다. 하지만 위에서 살펴봤듯이 screens는 종속변수 aud를 설명하는 주요한 변수이기에 제외하는 것이 바람직하지 못하다고 판단했다. 더불어 다른 변수로 변환하기도 여의치 않았다. 따라서 다음과 같이 교호작용 term을 고려한 모델을 만들어보았다.

Aud = screens year screens*year imdb_score facebook_likes budget * screens season genres rating

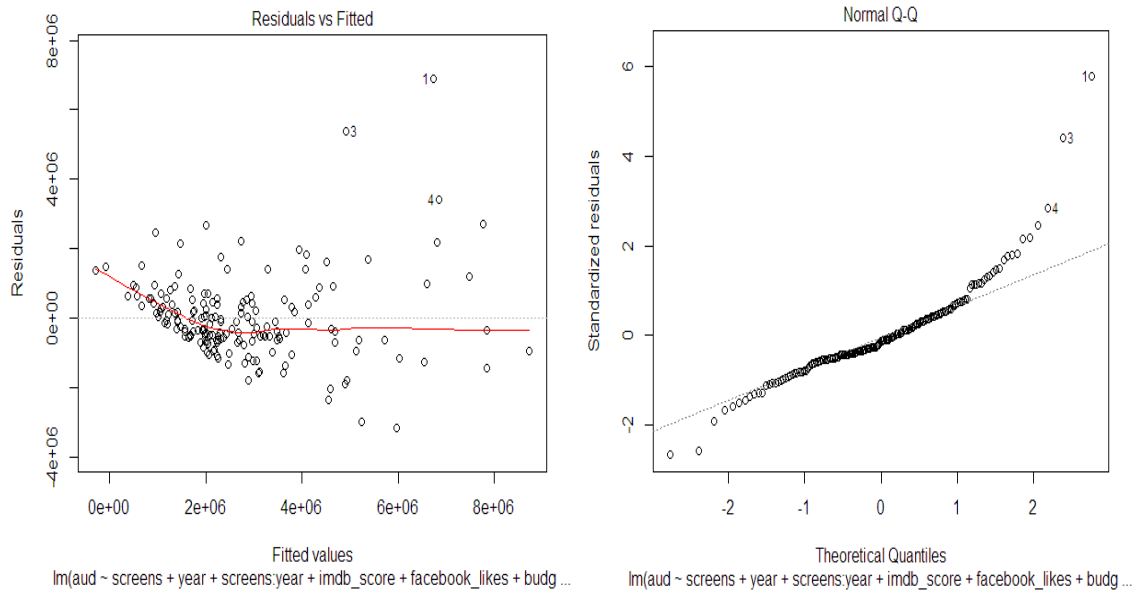
screens*year, budget*screens는 각각 screens + year + screens:year와 budget + screens budget:screens를 의미한다. ':'로 이어진 항은 연결된 두 변수의 교호작용을 나타낸다. 이는 높은 상관관계를 보이는 변수의 상호작용을 모형에 포함시켜 그 작용을 통제하기 위함이다. 비록 다중공선성에 대한 본질적인 해결이 되지 못하더라도 어느 정도 모형의 왜곡을 방지하는 효과가 있을 것이라 판단했다.

```
lm(formula = aud ~ screens + year + screens:year + imdb_score +
    facebook_likes + budget + budget:screens + season + genres +
    rating, data = mv)

Residuals:
     Min       1Q   Median       3Q      Max
-3151473  -623946  -164084   516599  6885157

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)    -1.505e+08  2.029e+08  -0.742  0.45939
screens         1.524e+06  2.849e+05   5.350 2.92e-07 ***
year            7.250e+04  1.010e+05   0.718  0.47388
imdb_score      5.908e+05  1.111e+05   5.317 3.40e-07 ***
facebook_likes  3.379e+00  4.873e+00   0.693  0.48901
budget         -1.129e-02  3.843e-03  -2.937  0.00379 **
season          3.855e+05  2.005e+05   1.922  0.05629 .
genres          1.232e+05  2.195e+05   0.561  0.57547
rating         -2.695e+05  3.167e+05  -0.851  0.39602
screens:year   -7.548e+02  1.417e+02  -5.328 3.24e-07 ***
screens:budget  7.958e-06  5.428e-06   1.466  0.14453
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1252000 on 164 degrees of freedom
Multiple R-squared:  0.6513,    Adjusted R-squared:   0.63
F-statistic: 30.63 on 10 and 164 DF,  p-value: < 2.2e-16
```

상기의 회귀 식을 분석한 결과이다. 교호작용 term이 screens: year과 screens: budget으로 나타남을 확인할 수 있다. 전체적인 모형의 설명력은 하단의 Adjusted R-squared로 파악 가능하다. 0.63으로 횡단면 자료임에도 상당히 높은 수치를 보여준다. F 통계량의 p-value 역시 매우 작은 수준이어서 모형의 설명력이 유의미함을 나타낸다.

개별적으로 변수를 살펴보자. 우선 변수의 부호 측면에서 대부분의 회귀계수가 예상한 바와 같은 부호를 갖는다. 예상외로 budget이 aud와 음의 상관관계를 보인다. 다음으로 변수의 유의성 측면에서 결과가 만족스럽지 못하다. P-value가 높아 유의성을 상실한 몇몇 변수가 눈에 띄기 때문이다. 교호작용 term이 유의미하여 생략하기 힘든 year 변수를 차치하고서라도, facebook_likes, genres, rating 그리고 screens:budget 변수의 유의성이 충분하지 못했다.

문제의 해결을 위해 그래프를 이용한 회귀 진단을 시행하였다. 회귀 분석의 기본 가정 중 관측치 간 독립성이 만족되었다고 가정했고, 등분산성과 정규성에 대한 진단을 각각 진행하였다.



좌측 그래프는 모형이 예측한 종속 변수 별 잔차의 분포를 보여준다. 외견 상 쉽게 알 수 있듯이 변수가 커질수록 잔차의 산포도가 커진다. 이는 전형적인 이분산의 특성이며 모형이 등분산성을 만족하지 못함을 뜻한다. 우측 그래프는 정규화 잔차의 분포를 보여준다. 직선을 따라 분포할 때 정규성 가정이 만족됨을 뜻하는데 비교적 직선과 일치하나 양쪽 꼬리가 긴 분포를 보이고 있다.

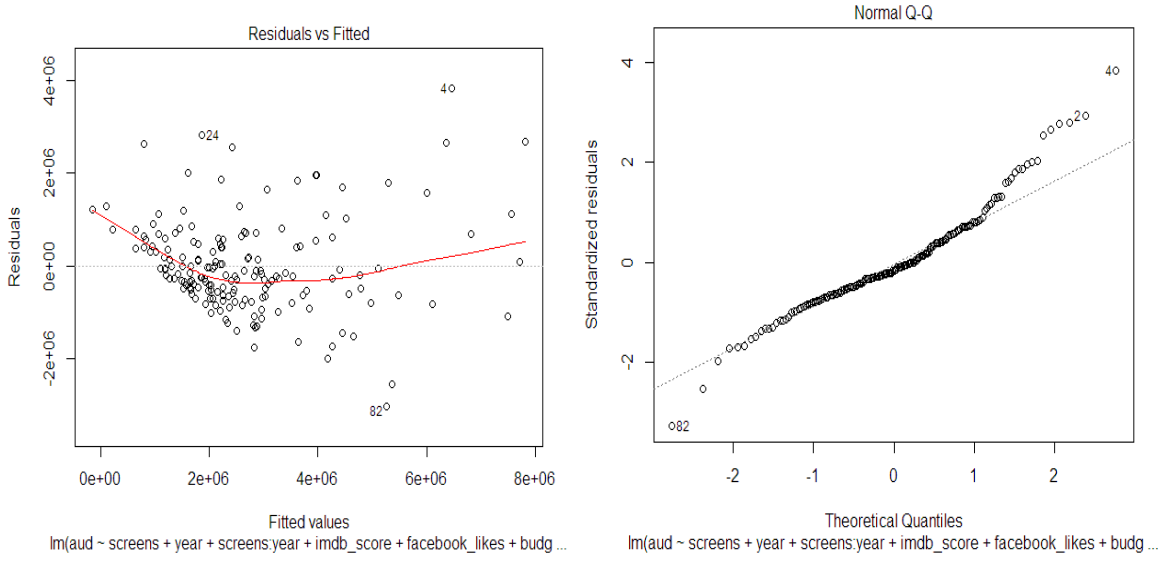
두 그래프는 아웃라이어의 존재도 드러낸다. 그래프에 표시된 숫자는 1, 3, 4번째 관측치에서 잔차 값이 비정상적으로 커지게 됨을 보여준다. 다시 말해 1, 3, 4번째 관측치는 모형의 예측에서 크게 벗어난 값이며, 이를 데이터에서 제외하면 모형의 적합도가 상승한다.

그러나 아웃라이어가 존재한다고 해서 무작정 제외시킬 수는 없다. 자의적인 관측치의 누락 역시 모형의 왜곡을 유발하기 때문이다. 모형의 적합성과 데이터의 완전성은 종종

상충하는 문제이며 이를 해결하기 위해 신중한 접근이 요구된다. 동분산과 정규성 가정에서 크게 벗어난 관측치 1, 3은 영화 Avatar와 Frozen(겨울왕국)이다. 두 영화는 각각 13,624,328명, 10,296,101명의 기록적인 흥행 스코어를 남겼다. Aud 변수의 평균값이 2,792,156임을 고려할 때 이 자체만으로도 특이 값이라 할 수 있다. 게다가 Avatar는 당시 전례가 없었던 3D 영화의 흥행이라는 파장을 낳았으며, Frozen은 애니메이션이라는 장르적 한계를 초월하여 대중의 사랑을 골고루 받았던 기념비적인 작품이라 할 수 있다.

	title	aud	screens	year	imdb_scor	facebook_	budget	season	genres	rating	country
1	title	aud	screens	year	imdb_scor	facebook_	budget	season	genres	rating	country
2	Avatar	13624328	912	2009	7.9	4834	2.37E+08	1	1	0	1
3	Avengers:	10494499	1843	2015	7.5	92000	2.50E+08	0	1	0	1
4	Frozen	10296101	1010	2013	7.6	2582	1.50E+08	1	0	0	1

비슷한 관객수를 기록했던 Avengers: Age of Ultron과 비교해보면 두 영화의 특이점은 더욱 두드러진다. 우선 screens 변수를 살펴보면 Avatar가 912, Frozen이 1010으로 1843의 스크린을 확보했던 Avengers의 절반 정도 밖에 미치지 못한다. 또한 유난히 높은 수치의 facebook_likes를 기록한 Avengers였음을 감안하더라도 두 영화의 facebook_likes는 지나치게 낮다. Facebook_likes의 평균 값인 22148에도 한참 모자란 수치인 것이다. 이는 배우 인지도 등 상업성의 측면에서 높은 평가를 받지 못하였던 두 영화가 이례적인 흥행 스코어를 기록했음을 의미한다. 이러한 이례적인 흥행의 조짐은 사전에 감지하기 매우 어려우며, 회귀분석을 통한 예측 모형에 적합하지 않은 사례임을 뜻한다. 따라서 관측치 1과 3은 모형을 왜곡하는 아웃라이어로 판단하였고, 모형에서 제외하였다.



관측치 2개를 제외한 후 나머지 173개 데이터에 대한 회귀진단 결과이다. 등분산성을

만족한다고 말하기 힘들지만 잔차의 산포 정도가 감소한 효과를 보이며 정규성 가정에도 더욱 잘 들어맞는다.

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    -2.094e+07  1.677e+08  -0.125  0.90078
screens         1.132e+06  2.390e+05   4.736 4.73e-06 ***
year            8.513e+03  8.346e+04   0.102  0.91889
imdb_score      4.887e+05  9.204e+04   5.310 3.57e-07 ***
facebook_likes  9.159e+00  4.056e+00   2.258  0.02526 *
budget         -1.039e-02  3.166e-03  -3.281  0.00127 **
season          2.972e+05  1.652e+05   1.799  0.07381 .
genres          2.051e+05  1.820e+05   1.127  0.26123
rating         -2.654e+05  2.603e+05  -1.020  0.30942
screens:year   -5.600e+02  1.188e+02  -4.712 5.26e-06 ***
screens:budget  6.002e-06  4.487e-06   1.338  0.18292
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1029000 on 162 degrees of freedom
Multiple R-squared:  0.6946,    Adjusted R-squared:  0.6757
F-statistic: 36.84 on 10 and 162 DF,  p-value: < 2.2e-16
```

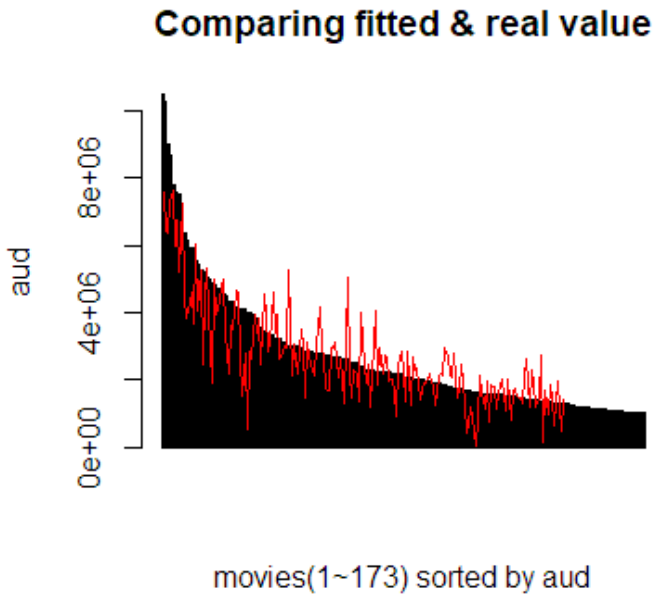
173개 관측치를 이용하여 기존 모형의 적합성과 유의성을 분석해보았다. 수정된 R-square 값이 0.63에서 0.6757로 상승했고 F 통계량 역시 증가하여 모형의 적합도가 향상되었다. 또한 개별 회귀변수들의 유의성이 높아져 변수 facebook_likes가 유의미한 변수가 되었다. 이는 모형 설정 단계에서 주요한 역할을 할 것으로 판단했던 배우 인지도의 유의미함을 확인한 결과로 그 의의가 있다.

이어서 변수 선택을 위해 step 함수를 이용하였다. direction = ‘both’ 구문을 이용하여 변수의 입출력을 반복적으로 진행하였다. AIC를 기준으로 하여 도출된 최적 모형은

aud = screens + year + imdb_score + facebook_likes + budget +season +screens:year 이다.

```
lm(formula = aud ~ screens + year + screens:year + imdb_score +
    facebook_likes + budget + season, data = out_mv)

Residuals:
     Min       1Q   Median       3Q      Max
-2817597  -604526  -185419   507505  3785388

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     8.187e+07  1.472e+08   0.556 0.578854
screens         1.038e+06  2.225e+05   4.667 6.31e-06 ***
year           -4.284e+04  7.318e+04  -0.585 0.559061
imdb_score      4.646e+05  9.028e+04   5.146 7.48e-07 ***
facebook_likes  9.455e+00  4.040e+00   2.340 0.020455 *
budget         -6.176e-03  1.589e-03  -3.888 0.000146 ***
season          2.205e+05  1.588e+05   1.388 0.166923
screens:year   -5.128e+02  1.105e+02  -4.641 7.03e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1030000 on 165 degrees of freedom
Multiple R-squared:  0.6883,    Adjusted R-squared:  0.6751
F-statistic: 52.05 on 7 and 165 DF,  p-value: < 2.2e-16
```

불필요한 변수가 제거된 결과 모형의 설득력이 증가하였다. 이를 최종 모형이라 할 수 있다. 총 독립변수는 6개에 추가로 screens와 year의 교호작용이 고려되었다. 다른 변인이 통제되었을 때, 독립변수 중 스크린 수, 작품성, 배우 인지도, 여름/겨울철 여부는 종속변수 관객수를 증가시킨다. 반면, 상영연도, 제작비는 오히려 관객수를 감소시킨다. 추가적으로, 스크린 수가 연도에 따라 증가하는 상관관계가 뚜렷하기에, 스크린과 연도를 곱해준 항이 스크린의 과대 평가된 영향력을 어느 정도 상쇄한다.



5. Expected original contribution

도출된 모형을 가지고 17년 개봉작 관객수를 95% 신뢰구간에서 추정해보았다.

```
                            fit       lwr     upr real_aud
Beauty and the beast    4517620 3609584.0 5425657  5138195
The mummy               1281823  287392.8 2276253  3661191
The Fate of the Furious 3408043 2579548.2 4236537  3653238
Thor: Ragnarok          4341015 3590868.0 5091162       NA
```

차례로 상반기 개봉했던 미녀와 야수, 미이라, 분노의 질주(상반기 관객수 top3) 그리고 17년 11월 현재 상영중인 토르: 라그나로크 이다. 미이라를 제외하고는 추정치가 실제 값에 부합하는 결과를 보인다. 또한 토르: 라그나로크의 경우, 현재 상영중인 영화의 최종 관객수 추정이 가능함을 의미한다. 더 나아가, 독립변수로 쓰일 값의 합리적인 예측이 가능하다면 제작 단계 혹은 기획 단계의 영화 흥행 여부 추정 역시 어렵지 않을 것이다.

6. Limitation and consideration for future research

하지만 모형의 예측을 그대로 수용하는 것에는 문제가 있다. 앞서 언급했듯이 회귀진단 결과 데이터의 분포가 일정하지 않아 동분산을 가정할 수 없기 때문이다.

```
> bptest(out_m)

        studentized Breusch-Pagan test

data:  out_m
BP = 34.407, df = 10, p-value = 0.0001575
```

위의 결과와 같이 이분산은 BP Test에서 더욱 잘 드러난다. 그렇기에 회귀모형의은 결함이 있으며, 모형의 사용에 신중한 접근이 요구된다. 구체적으로, 이분산이 확인되었기에 일반최소자승법(GLS)나 가중최소자승법(WLS)와 같은 방법으로 모형을 수정할 수 있을 것이다. 또한 모형 도출 과정에서 데이터가 독립적이라 가정되었지만 이는 현실과 다를 수 있다. 개별 영화의 관객수는 그 영화의 데이터로만 결정되지 않기 때문이다. 예를 들면, 흥행작과 비슷한 시기에 상영된 영화는 모형의 예측치보다 적은 관객 스코어를 기록할 가능성이 높다. 영화 간 독립적 아니라 대체적 관계가 형성되는 것이다. 그렇기에 모형은 자기 상관의 위험이 있다. 이 문제를 해결하기는 쉽지 않기에 시간 개념의 도입 등 더 고차원적인 노력이 요구된다.

그럼에도 회귀 모형의 결과는 앞서 제시한 독립변수와 종속변수 간의 상관관계를 확인하였다는 측면에서 의의를 지닌다. 대체로 모형의 설명은 사전에 예측된 변수의 영향력

과 크게 동떨어지지 않았다. 때문에 분석의 주된 목적이었던 독립변수를 이용한 종속변수의 예측 가능성은 여전히 유효하다 할 수 있을 것이다.

이 밖에도 다중 공선성의 문제가 크다. 6페이지에서 제시한 상관계수 매트릭스를 보면 독립변수 간의 상관관계가 무시할만한 수준이 아님을 알 수 있다. 이는 특히 screens 변수의 존재로 인해 두드러지는데, 다른 변수와의 상관관계가 특히 높아 다중공선성이 의심된다. 하지만 screens 변수를 제외하였을 때 모델의 설명력이 너무 낮아 일단은 포함하였다. Ridged Regression이나 PCA 등의 다른 방법을 강구할 필요가 있다.

7. Reference

http://www.imdb.com/

http://www.kmdb.or.kr/

한국 영화시장의 흥행결정 요인에 관한 연구

- 2006-2008년 개봉작품을 중심으로_박승현, 정완규(2009)

한국 영화의 흥행성과 결정요인 분석:

-2012년 개봉 상업영화를 중심으로_김상호, 한진만(2014)