추**1.** **: 분석 모형의 개요**

1-1. 주제 선정 동기: 2018년 기준 1인당 연간 영화 관란횟수는 4.18회로 미국의 3.7회보다 높은 수치.[[1]](#footnote-0) 즉 영화 산업이 다른 나라에 비해 큰 편이라는 것인데, 과연 매달, 다양하게 개봉되는 영화 중 흥행하는 영화는 어떻게 결정되고, 관객들이 그 영화를 고르는데에 어떤 요소들이 영향을 미쳤을지에 의문을 갖게 되었다.

1-2. 목적 : 영화 흥행에 영향을 미치는 요소들이, 흥행의 지표라고 판단되는 매출액에 어느정도 상대적인 영향력을 미치는지 분석해보고자 하였다.

더 나아가, 향후 개봉영화 흥행 여부를 예측하기 위해, 매출액에 영향을 미치는 주요 요소에 대해 알아보고, 기대 매출액 달성을 위해 필요한 투입량을 예상해 보고자 한다.

1-3. 모형 설정과정

­­“영화의 흥행요소는 창조의 영역, 배급의 영역, 홍보의 영역, 경쟁의 영역으로 분류가 가능한다” [[1]](https://d.docs.live.net/11f30f74b3d3639d/%EB%AC%B8%EC%84%9C/%EA%B3%84%EB%9F%89%EA%B2%BD%EC%A0%9C%ED%95%99%20%EB%B0%9C%ED%91%9C%20%EA%B3%84%ED%9A%8D%EC%84%9C.docx#_ftnref1) 전희국, 『영화 흥행 실적 예측을 위한 빅데이터 전처리 (Big Data Preprocessing for Predicting Box Office Success)』, KIISE, 제20권 제12호(2014.12)

문헌자료를 참고하여 각 영역에 해당하는 독립변수를 설정하였다.

(이에 대해 자세한 얘기는 추정모형 설명할 때 더 자세히)

**2.** **추정모형**

2-1. 변수설정.

- 종속변수 Y: log(매출액)

가장 바람직한 흥행지표는 비율이라고 생각되지만 투자액은 투자사의 대외비로 분류되어 있어 정보의 접근이 어렵다. 차선책으로 관람객수영화티켓비용을 의미하는 매출액을 종속변수로 삼는다.

탑 200 누적매출액 순

y변수의 분산 및 단위가 너무 크고,

독립변수 1단위 변화에 따른 종속변수 상대적인 변화 나타낼 거임

- 독립변수 X

: 영화 흥행에는 참신함(Creativeness), 분배정도(Distribution), 화제성 및 대중성(Marketing), 경쟁력(Competition) 등이 영향을 미친다고 가정한다.[[1]](https://d.docs.live.net/11f30f74b3d3639d/%EB%AC%B8%EC%84%9C/%EA%B3%84%EB%9F%89%EA%B2%BD%EC%A0%9C%ED%95%99%20%EB%B0%9C%ED%91%9C%20%EA%B3%84%ED%9A%8D%EC%84%9C.docx#_ftn1)

② : 로튼 토마토 지수 → 참신함을 위한 지표

② : 스크린 수/100 → 분배정도 (스크린(코비스)

해당 영화 상영 기간 중 최대 스크린수, 전국 영화 전체 스크린수

③ : 인스타그램 해시태그 수/1000000 → 화제성

④ : 네이버 관람객 평점 → 대중성

⑤ : 개봉 달 동시상영 영화개수 → 경쟁력

- 가변수 D

: 책 또는 만화로 원작이 존재할 경우, 원작의 팬 층이 영화 매출에 영향을 줄 것이다. 따라서 가변수를 통해 원작 존재여부(r=1: 존재, r=0: 존재하지 않음)에 따른 절편 변화를 알아본다.

로튼 토마토 지수란?

: **로튼 토마토**(Rotten Tomatoes)는 [영화](https://ko.wikipedia.org/wiki/%EC%98%81%ED%99%94) 관련 [웹사이트](https://ko.wikipedia.org/wiki/%EC%9B%B9%EC%82%AC%EC%9D%B4%ED%8A%B8) 가운데 하나이다. 영화에 대한 소식, 비평, 정보 등을 제공한다. 주로 비평가 위주의 평점을 매기는 곳이다. 생소하실 수도 있지만, 가끔 포스터에서 ‘로튼토마토 지수 100%’라는 문구를 본적이 있을 것이다.

(출처 위키백과이긴한데…ㅜ)

2-2. 방향성 예측 및 가설

-

경쟁은 매출액과 음의 관계, 그 외에는 양의 관계를 가질 것이다. 특히 스크린수가 가장 큰 영향을 줄 것으로 예측한다.

**3.** **사용 자료 (data)**

- 데이터는 www.kobis.or.kr 에 공개된 자료에서, 누적매출액이 높은 200개의 영화를 선정하였다.

- 데이터 수집 주소 : [www.kobis.or.kr](http://www.kobis.or.kr/) / [www.rottentomatoes.com](http://www.rottentomatoes.com/) / [www.instagram.com](http://www.instagram.com/) / movie.naver.com

자료의 기초적인 분석

| N=200 | 매출액 | 스크린수 | 네이버 평점 | 개봉한영화 | 해시태그 수 | 로튼토마토 | 원작여부 |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 평균 | 4.71E+10 | 1064.175 | 8.5298 | 2.01 | 197899.555 | 50.585 | 0.495 |
| 표준 편차 | 2.36E+10 | 487.2114 | 0.739446323 | 0.913274522 | 1213714.884 | 38.79546644 | 0.50123 |
| 최 소값 | 2.06E+10 | 238 | 5.58 | 1 | 18 | 0 | 0 |
| 최 대 값 | 1.4E+11 | 2835 | 10 | 4 | 12721432 | 100 | 1 |

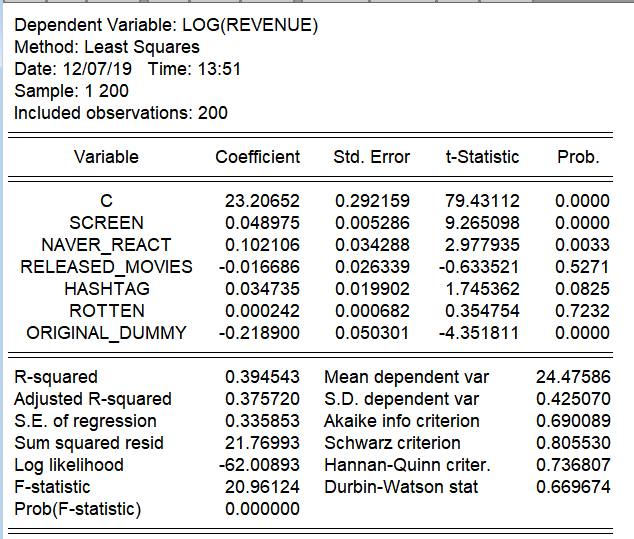
평균, 최대값 최소값 단위 맞춰서 가볍게 말하면 될 듯

변수 간의 상관관계

|  | 매출액 | 스크린수 | 네이버 평점 | 같은 달에 개봉한 영화개수 | 인스타그램 해시태그 수 | 로튼토마토 | 원작여부 |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 매출액 | 1.00 |  |  |  |  |  |  |
| 스크린수 | 0.51 | 1.00 |  |  |  |  |  |
| 네이버 평점 | 0.26 | 0.17 | 1.00 |  |  |  |  |
| 같은 달에 개봉한 영화개수 | 0.04 | 0.04 | 0.01 | 1.00 |  |  |  |
| 인스타그램 해시태그 수 | 0.07 | -0.04 | -0.12 | -0.05 | 1.00 |  |  |
| 로튼토마토 | 0.19 | 0.32 | 0.30 | -0.02 | -0.06 | 1.00 |  |
| 원작여부 | -0.14 | 0.23 | -0.02 | -0.11 | -0.10 | 0.21 | 1.00 |

매출액이랑 어떤변수랑 음/양의 상관관계를 가진다?

**4. 추정방법**

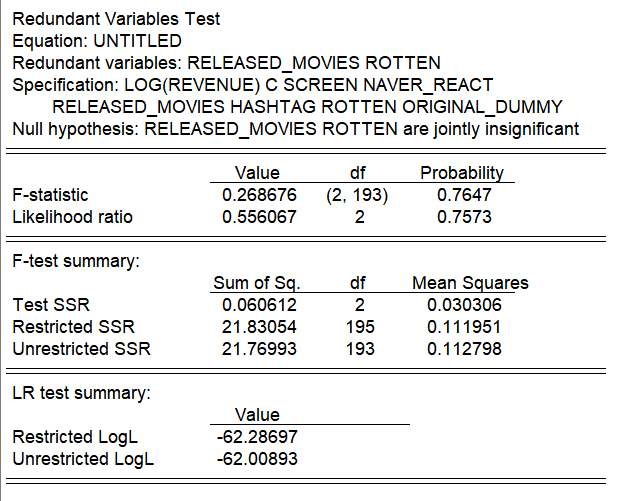
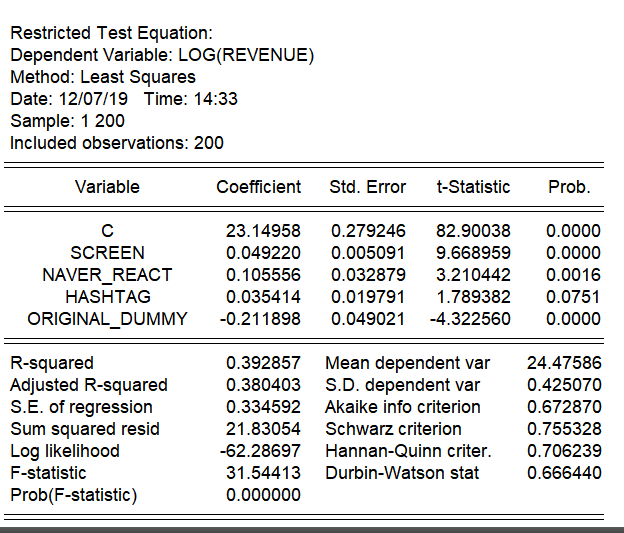
1. 설정모형에 대한 회귀분석

* 1. screen수, 네이버 평점, 원작여부가 5% 유의수준에서 유의한 변수로 나타났다.
* 인스타그램 해쉬태그도 5% 유의수준에서는 기각되지만, 10% 수준으로 높일 경우 유의미한 변수로 취급된다..

3. 같은 달에 개봉한 영화의 수를 나타내는 Released\_movies와 로튼토마토 평가를 나타내는 Rotten변수의 p-value가 각각 0.53 과 0.73으로 높은 값을 보이며 10% 5% 유의검정에서 모두 유의적이지 않은 것으로 나타났다

1. 부적합한 변수의 제거

-



초기의 모형에서 released\_movies와 rotten에 대해 부적합 변수에 대한 검정을 시행한 결과 위와 같이 나타났다.

먼저, F-statistic의 유의확률을 나타내는 probaility가 5%, 10%의 유의검적에서 모두 기각되지 않기때문에,

released\_movies와 rotten변수는 유의미하지 않다 라는 가정은 기각되지 않는다.

따라서, released\_movies와 rotten변수는 유의미하지 않다.

정보 기준(information criteria)을 고려해 자료를 살펴보면, AIC(Akaike infromation criterion) 값이0.6901 에서 0.6728 로, SC(Schwarz criterion) 값이 0.8055에서 0.7553로 하락하는 것을 관찰할 수 있다. 따라서 변수 released\_movies와 rotten을 제외한 모형이 더 적합한 모형임을 알 수 있다.

또한 나머지 변수 screen, naver\_react, hastag, orignial\_dummy는 각각 유의확률 0.0000, 0.0016, 0.0751의 값을 보이면서 10% 유의수준에서 변수가 유의적이지 않다는 귀무가설을 모두 기각할 수 있다. 따라서 위의 네 변수는 모두 유의적인 변수임을 알 수 있다.

따라서 다음과 같은 새로운 회귀분석 식을 도출하였다.

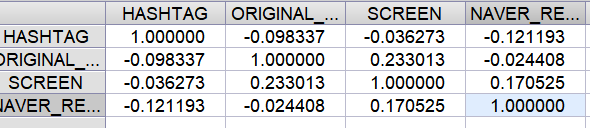
log(Revenue) = c + b1screen +b2naver\_React+b3hastag+b4original\_dummy +ei

1. 다중공선성 점검

-다중공선성의 점검은 모형의 추정 결과에 있어서 높은 을 보이면서 동시에 개별 회귀계수에 대한 낮은 t-검정치가 나타나는 경우 의심해볼 수 있다.

(우리는 r^2이 엄청 큰게 아니라서 위에 있는 건 못쓸듯)

-설명변수들 사이의 상관계수 직접 계산하여 이 값이 모형 전체의 R^2보다 큰경우

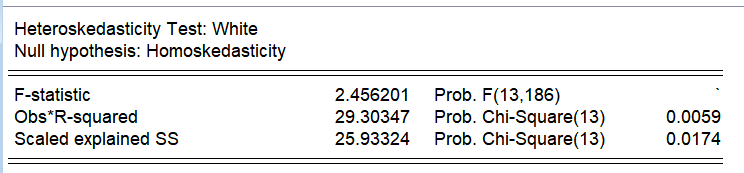


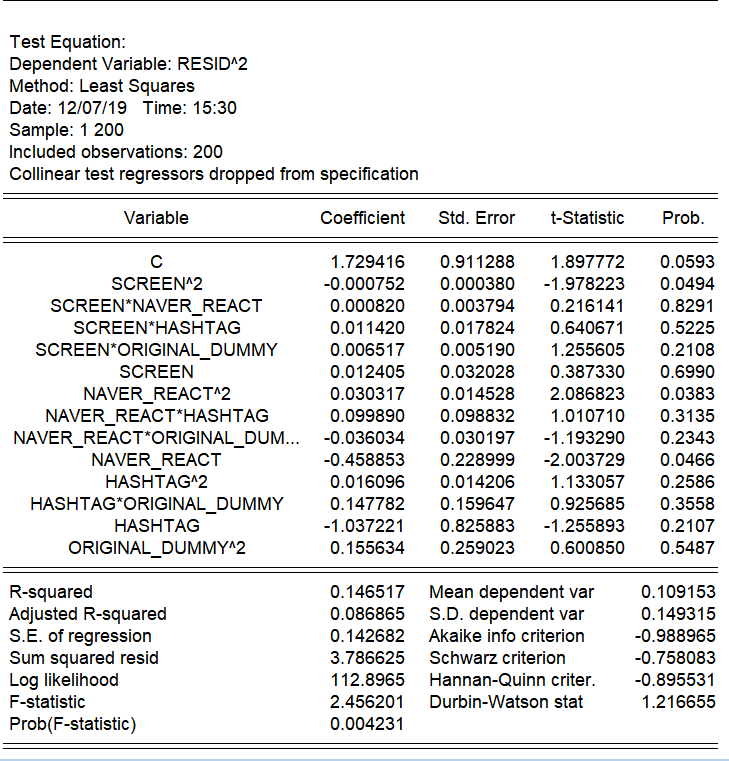
설명변수 사이의 상관계수 중 모형 전체의 R^2인 0.3928 보다 높은 변수는 없어서

위 회귀식에서 다중공선성의 문제는 나타나지 않음을 알 수 있다.

1. 이분산 검정

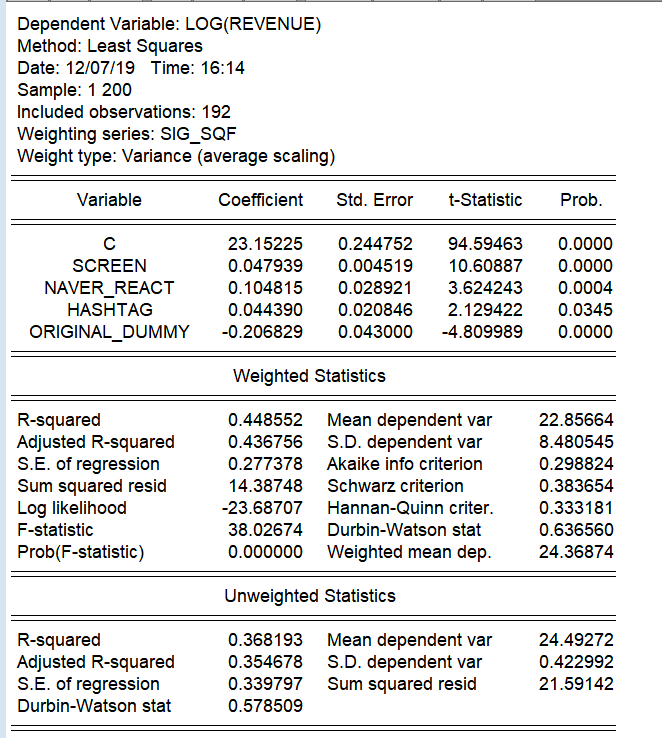
- white의 이분산 검정법이용해서 이분산 존재여부 판단해 봄





White의 이분산 검정법을 적용해 본 결과, 검정통계량 n\*R^2 = 29.30347및 이에 대한 유의확률은 0.0059가 도출되었다. 5% 유의수준 하에서 검정통계량이 임계치인 카이스퀘어(13;5%) = 22.362보다 큰 값을 보이고, 유의확률 또한 0.05보다 작은 값을 보이므로 이분산이 없다는 귀무가설이 기각됨을 알 수 있다.

따라서 설정한 모형에 이분산이 있다는 결과를 얻을 수 있다.



이분산 해결 방법: 일반최소 자승법(FGLS )

FGLS를 적용한 결과, 조정된 결정계수 값이 = 0.4368의 값이 나타나 [자료 2](log(Revenue) = c + b1screen +b2naver\_React+b3hastag+b4original\_dummy +ei

)의 조정된 결정계수 값 = 0.3757 에 비해 높아진 것을 확인할 수 있다. 또한 각 변수들의 p-value 값도 [자료 2]에 비해 모두 더 낮은 값을 보인다.

**5. 자기상관**

Durbin Watson 통계량이 2에 수렴하고 있지 않아, 자기상관이 존재할 가능성은 있지만,

이러한 자기상관이 발생한 이유는, 누락변수때문일 확률이 높다.

누락변수를 찾아 자기상관을 해결하기 어려우므로 이는 본 연구의 한계점이라고 할 수 있다.

**5. 주요 추정결과**

**부적합변수 판정을 통해 초기에 설정한 설명변수들 중 released\_movies, rotten 변수를 제외한**

log(Revenue) = c + b1screen +b2naver\_React+b3hastag+b4original\_dummy +ei

**의 최종 회귀식을 도출해냈다. 위 회귀식에서 다중공선성의 문제는 나타나지 않았지만, 이분산이 존재하는 것으로 확인되어 FGLS를 적용해 회귀분석한 결과는 [자료 7]과 같다. 이를 통해 도출된 회귀계수를 대입하면 다음과 같은 추정치를 도출할 수 있다.**

**log(revenue)= 23.15225+ 0.047939\*Screen + 0.104815\*Naver\_react + 0.044390\*Hashtag -0.206829\*Original\_dummy**

**밑에 t통계량 써서 예쁘게 만들기**

1. **screen수가 100개 더 증가함에따라 매출액은 4.8% 증가한다.**
2. **naver 평점이 1점 증가하면 매출액은 10.5%증가한다.**
3. **인스타그램 해시태그가 100만개 증가하면 매출액은 4.4% 더 증가한다.**
4. **원작이 있으면, 매출액이 21% 떨어진다. (1편 또는 원작의 기대치를 속편이 충족시키지 못했거나, 이미 스토리를 알고 있어서 사람들이 찾아보지 않았을 거라고 추측해봄)**

**→ 모든 변수(원작유무 제외)가 한 단위씩 증가하면, 매출이 4.8% + 10.5% + 4.4% = 19.7% 증가한다.**

**6. 영화 매출 예측**

**현 상영되고 있는 겨울왕국2는 2019.12.07 기준, 매출액이 1,997,674,070원.**

**현재 역대 최고 매출액은 극한직업의 139,651,845,516원이므로,**

**상영 종료 후 겨울왕국2가 역대 최고 매출액을 달성하기 위해서는,**

**네이버 평점: 8.92 → 9.92 : +10.5%**

**인스타그램: 207,478+288+6703 = 214,469 → 1,214,469 +4.4%**

**스크린수: 2,648 → 2,800 +7.296%**

**원작: 1 → 1**

**겨울왕국 2가 종영까지**

**screen 최대 장악수인 2800개를 도달하고**

**인스타 누적 121만개**

**네이버 평점이 9.92까지**

**를 달성한다면**

**매출액이 현재보다 22% 증가할 것이다.**

**새로운 영화예측 <월터의 상상은 현실이 된다>**

**네이버 평점 8.80 | 인스타그램 7,504 | 스크린 수 378 | 원작 1**

**예측 매출액 = log(Walter) = 24.0514 #자연로그**

**실제 매출액 = 7,001,955,999원 → log(real\_Walter) = 22.6695**

**<한계>**

1. 각 독립변수가 해당영역의 대표값이 되는데에 한계가 있다.

(예, 네이버 평점이 대중성을 완벽하게 대표하지는 못한다.)

2 rotten tomato 지수가 없는 영화도 있었다. 그럼에도 rotten 지수를 제외하지 않은 이유는, rotten지수가 국내외 전문가들의 의견이 반영된 자료이기 때문이다.

1. 영화진응위원회 영화정책연구원 「2018년 한국 영화산업 결산」, p.9-10 [↑](#footnote-ref-0)