“Simple is the Best”

- 예측모델을 사용한 영화 관객수 예측 -

cau_stat

데이터 분석 개요

분석 목표

1. 개봉예정 영화 관객수 예측

분석방법

1. 다중 회귀 모형
2. 랜덤 포레스트

데이터 수집 기간 및 채널

1. 기간: 2015.01.01.~2018.07.13.
2. 채널: 영화진흥위, 네이버 영화
3. 수집도구: Python, R

수집 데이터

1. 독립 변수 개수: 50개
2. 표본 개수: 520개

역할 분담

강동현

중앙대 응용통계학과
데이터 분석 (R)

이승화

중앙대 응용통계학과
데이터 수집 및 PPT

한승훈

중앙대 응용통계학과
데이터 분석(Python)

성의창

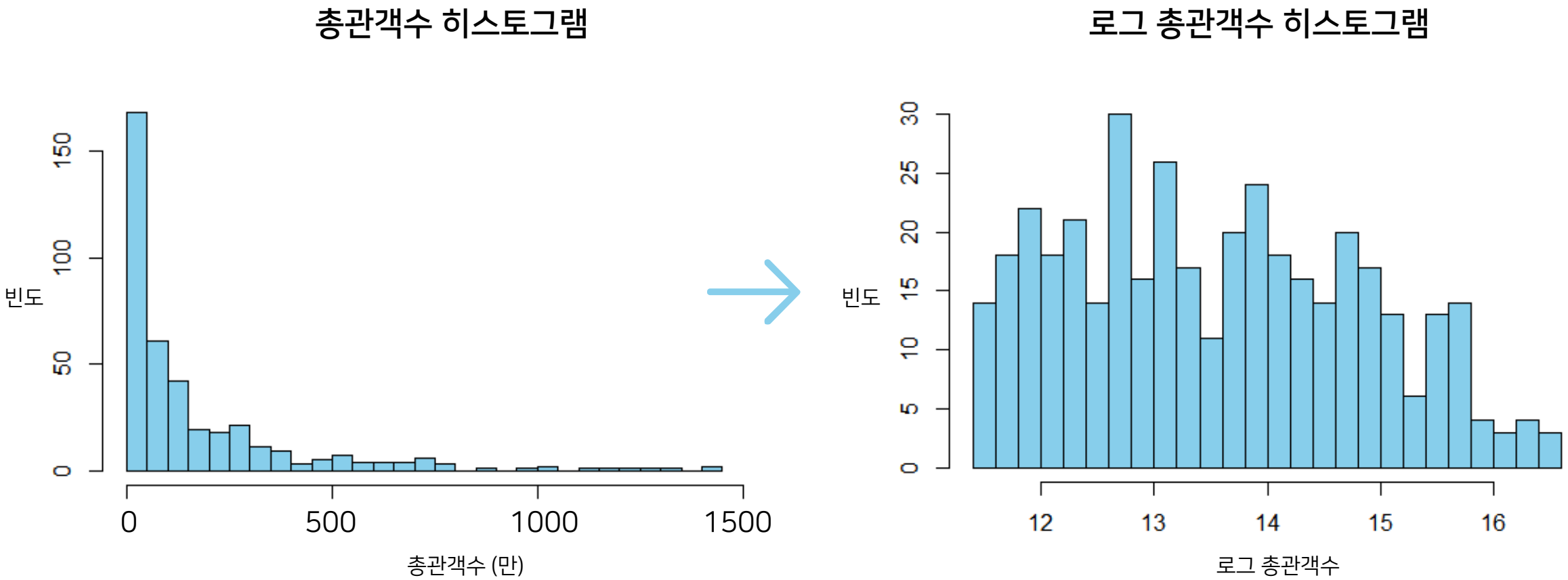
중앙대 응용통계학과
데이터 분석(Python)

이영훈(팀장)

중앙대 응용통계학과
데이터 분석(R)

목차

| 종속 변수



편향된 총관객수 분포를 로그변환을 통해 개선

| 내적 변수

장르

액션 → 1
멜로/로맨스 → 2
드라마 → 3
공포 → 4
(애니메이션의 경우 미포함)

배우

최근 3년 동안
주연배우의 관객수 평균

관람 등급

전체 관람가 → 1
12세 관람가 → 2
15세 관람가 → 3
청소년 관람불가 → 4

제작 국가

한국영화 → 1
그 외 → 0

| 외적 변수

관객수

종속 변수와 동일한 변환

스크린 수

종속 변수와 동일한 변환

평점

평점 → 스케일링
참여자수 → 종속 변수와 동일한 변환

기타

기대지수 → 종속 변수와 동일한 변환

| 가공 변수

경쟁 변수

영화 개봉일 전후 10일 이내에
개봉한 10만 이상의 영화 수

단독 개봉

특정 영화관에서만 상영 → 1
복수 영화관에서 상영 → 0

연휴 변수

연휴에 개봉 → 1
그 외에 개봉 → 0

지속 변수

제곱근 변환

| 가공변수 - 지속 변수

<23일 이후의 관객수 변동률과 상관성이 가장 높은 비율 찾기 위해 상관분석 실시 >

변수명	상관계수
Rate1	0.82
Rate2	0.86
Rate3	0.91

Rate1 : 18일부터 23일까지의 누적 관객수 변동률

Rate2 : 19일부터 23일까지의 누적 관객수 변동률

Rate3 : 22일부터 23일까지의 누적 관객수 변동률

상관분석 결과 Rate3이 23일 이후의 관객수 변동률과 가장 상관성이 높아
Rate3을 지속변수로서 채택

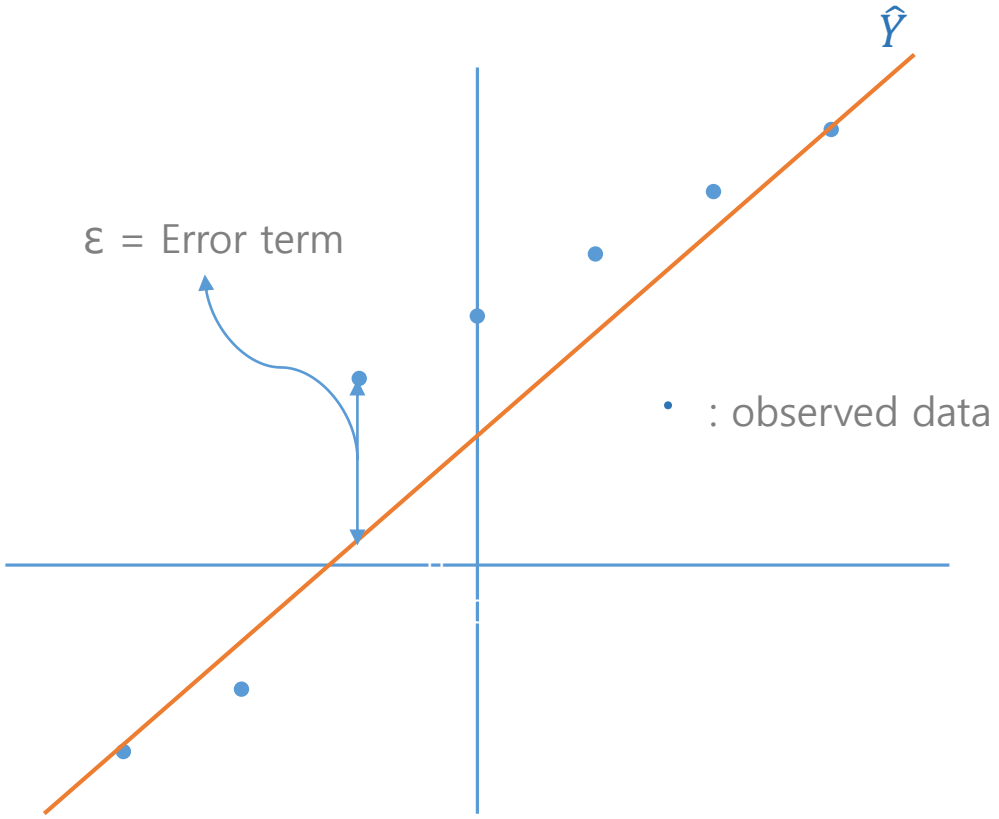
| Linear Regression

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

Predicted Values of Y

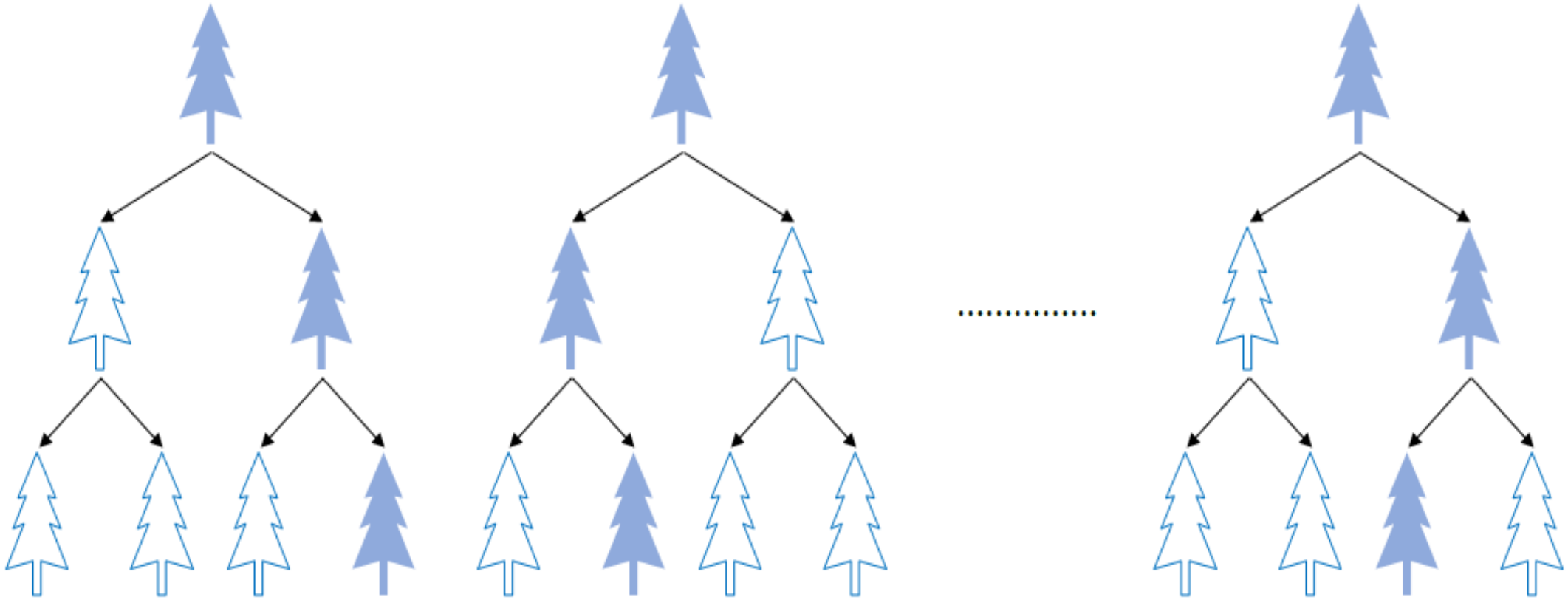
Y-intercept = Level of Y When x is 0

$\hat{\beta}_1$ = slope = rate of Predicted ↑/↓ for Y scores for each Unit increase in X



종속 변수와 한 개 이상의 독립 변수의 선형 상관 관계를 모델링하는 분석 방법

| Random Forest



의사결정나무들의 예측 결과들을 종합하여 추정하는 방법

| K -Fold Cross validation

k개의 fold를 만들어서 진행하는 교차검증

03. 분석 과정

| 2가지 모델

일	월	화	수	목	금	토
19	20	21		23	24	25
26	27	28	29	30	31	9/1
2	3	4	5	6	7	8
9	10	11		13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
30						

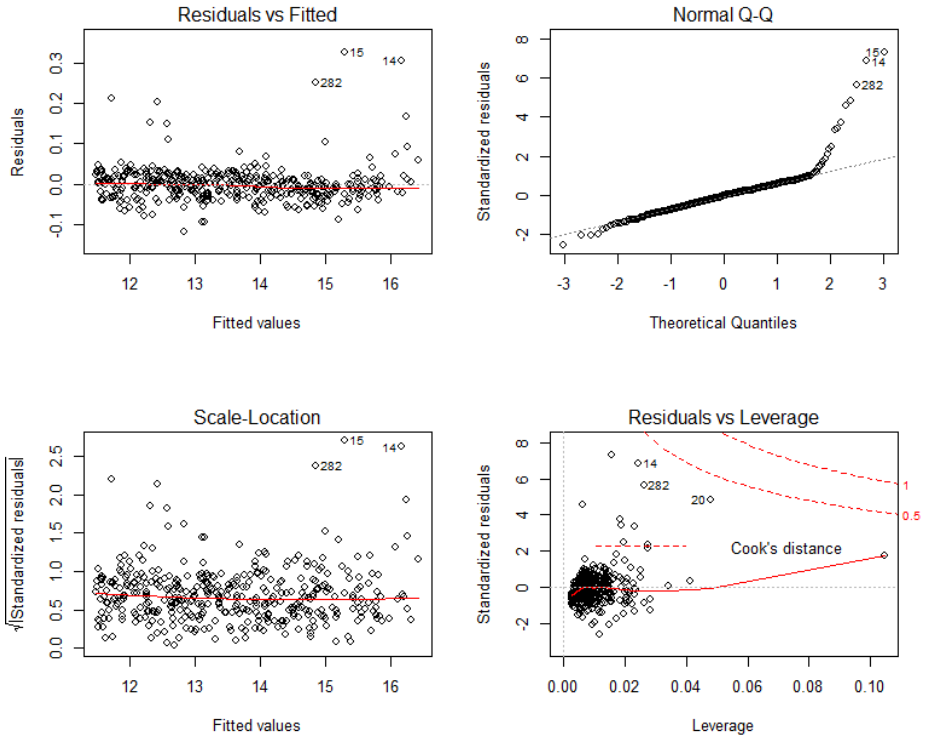
‘물괴’의 개봉일이 다른 두 영화의 개봉일과 다르기 때문에 두 개의 모델 설정

| 첫 번째 모델 – 선형 모형

$$\widehat{logY} = \hat{\beta}_0 + \hat{\beta}_1\sqrt{X_1} + \log X_2 + \log X_3$$

Y = 총관객수 X_1 = 지속변수 X_2 = 23일차 누적관객 수 X_3= 경쟁변수

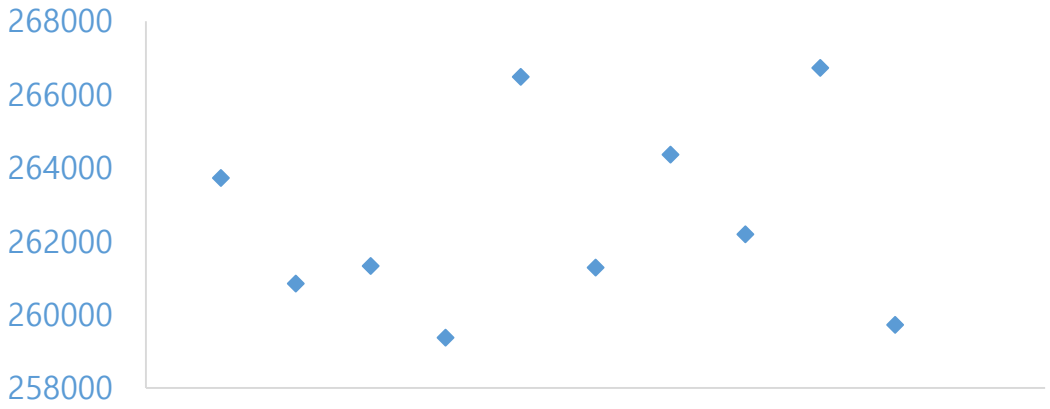
변수	계수	유의확률	분산 팽창 계수
절편	$\hat{\beta}_0$=0.092942	0.00649	
X_1	$\hat{\beta}_1$=1.7412	<2e-16	1.1614
X_2	$\hat{\beta}_2$= 0.9922	<2e-16	1.2673
X_3	$\hat{\beta}_3$=-0.0154	0.05015	1.1366
R^2 = 0.9988		모델유의확률 = <2.2e-16	



| 모형 선택

$$RMSE = \sqrt{\frac{1}{\text{총 영화데이터 수}} \sum (\text{실제 관객수} - \text{예측 관객수})^2}$$

LINEAR REGRESSION MODEL1 RMSE



262602

RANDOM FOREST MODEL1 RMSE

420275

Root mean square error(RMSE) 가 Linear model 이 더 작으므로 선택

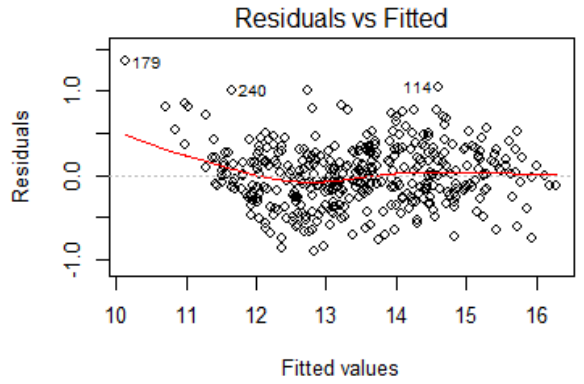
| 두 번째 모델 – 선형 모형

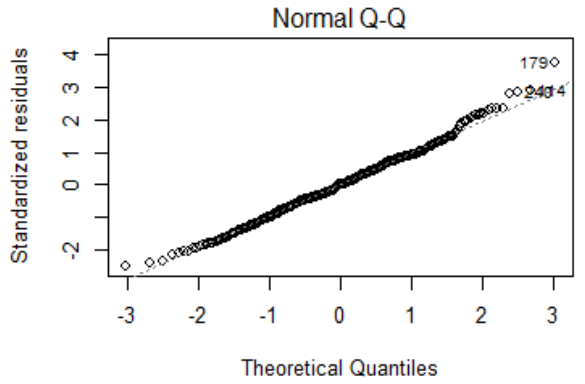
$$\widehat{logY} = \hat{\beta}_0 + \hat{\beta}_1 logX_1 + \hat{\beta}_2 logX_2 + \hat{\beta}_3 X_3 + \hat{\beta}_4 logX_4 + \hat{\beta}_5 X_5 + \hat{\beta}_6 logX_6 + \hat{\beta}_7 logX_7 + \hat{\beta}_8 logX_8 + \hat{\beta}_9 logX_9$$

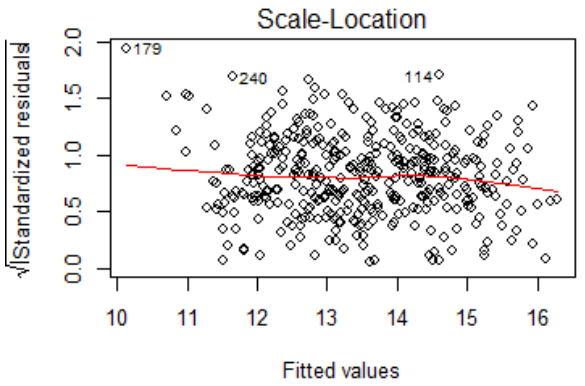
Y= 19일차 관객수 X_1 = 배우 X_2 = 2일차 박스오피스 순위 X_3= 연휴 X_4= 2일차 관객수 X_5=제작국가

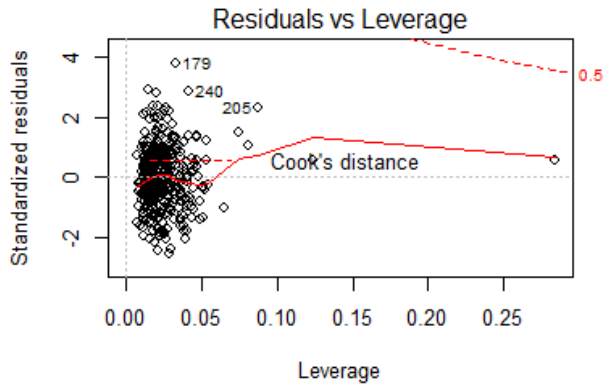
X_6 = 시사회 관객수 X_7 = 시사회 스크린 수 X_8 =개봉 전 좋아요 수 X_9 = 경쟁변수

변수	계수	유의확률	분산 팽창 계수
절편	$\hat{\beta}_0$= 3.2598	4.30e-10	
X_1	$\hat{\beta}_1$= 0.2195	4.38e-12	2.6946
X_2	$\hat{\beta}_2$=-0.4632	<2e-16	3.7120
X_3	$\hat{\beta}_3$=-0.0991	4.89e-06	1.1584
X_4	$\hat{\beta}_4$=-0.5747	<2e-16	5.2359
X_5	$\hat{\beta}_5$= 0.1279	9.29e-09	1.4199
X_6	$\hat{\beta}_6$= 0.1626	2.39e-08	5.3655
X_7	$\hat{\beta}_7$=-0.0984	8.21e-05	4.5641
X_8	$\hat{\beta}_8$= 0.0532	0.00793	1.7978
X_9	$\hat{\beta}_9$=-0.1154	0.07130	1.1542
R^2 = 0.9192		모델유의확률 = <2.2e-16	









| 모형 선택

$$RMSE = \sqrt{\frac{1}{\text{총 영화데이터 수}} \sum (\text{실제 관객수} - \text{예측 관객수})^2}$$

LINEAR REGRESSION MODEL2 RMSE

766513

RANDOM FOREST MODEL2 RMSE

680026

Root mean square error(RMSE) 가 Random forest 가 더 작으므로 선택

04. 결과 도출

| 영화 별 관객수 예측

2,918,481

271,007

1,250,791

THANK YOU
Q & A