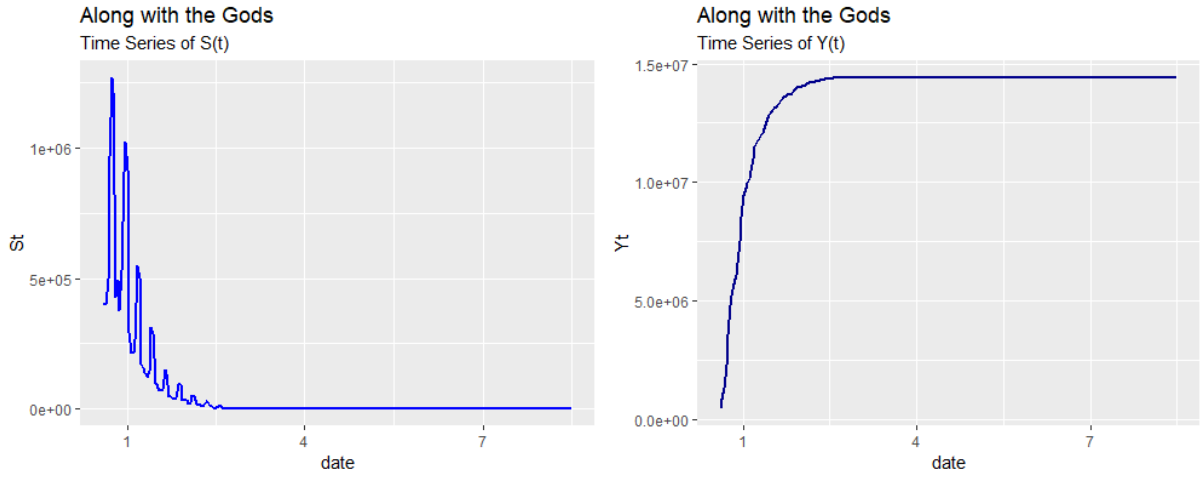
영화 흥행 예측

[영화 1] 신과 함께 - 죄와 벌 (2017.12.20 개봉)

1) 개봉 후 일별 관객수 S(t)와 누적 관객수 Y(t)의 시계열 도표



➔ 다음은 2017년 12월 20일에 개봉한 '신과 함께 - 죄와 벌'의 개봉 후 현재까지 일별 관객수(S(t)) 및 누적 관객수(Y(t)) 시계열 도표를 그린 것이다. 먼저 S(t)의 그래프를 살펴보면 초반에 급격하게 관객 수가 증가하였다가 감소함을 볼 수 있다. 그래프에서 유독 솟아 있는 부분은 주말의 일별 관객수 값이다. 평일에 비해 주말의 관객수가 많게는 거의 2배 정도였다. 따라서 관객 수 예측을 할 때 휴일효과를 보정해야 함을 알 수 있었다. 다음으로 Y(t)의 그래프를 살펴보면 누적 관객수 역시 초반에 급격하게 증가하였다가 어느 순간 일정해짐을 볼 수 있다. 이는 초반에 영화가 개봉한 후 궁금함, 입소문 등으로 많이 본 뒤 일정한 시간이 지나고 나서는 영화의 수요가 빠르게 준다는 것을 알 수 있다.

2) 확산 모형을 이용한 최적 예측 모형 결정

앞에서 그래프를 통해 휴일효과를 적절히 보정해야 함을 알 수 있었다. 따라서 기간 내의 공휴일, 주말(토요일, 일요일)은 일별 관객수를 반으로 나누어 2일로 간주하기로 했다. 예를 들어, 12월 24일의 일별 관객수는 1,268,537명이지만 일요일이므로 24일을 이틀로 생각하여, 일별 관객수를 반으로 나눠서 634,268.5명이라고 생각하였다.

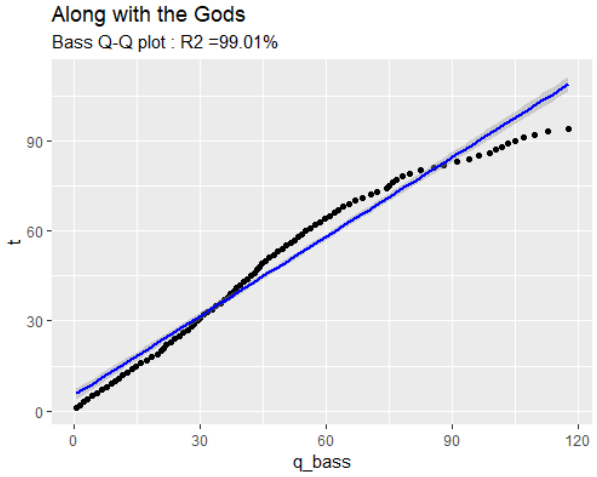
(ex) 2017-12-24 1268537 ➡ 2017-12-24 634268.5
2017-12-24 634268.5

데이터를 보정한 후 Bass, Gumbel, Logistic, Exponential의 4가지 확산 모형을 이용하여 총관객수를 예측해보고자 한다. 각 모형에 대해 OLS 추정법을 통해 1주차, 2주차, 4주차에 해당하는 7일, 14일, 28일까지의 데이터로 각각 총 관객수 m을 예측하였다.

오른쪽 표는 n과 모형의 분포에 따른 상대 오차 값을 나타낸 것이다. 상대오차는 $100 * (\hat{m} - m)/m$값으로 이 때의 m은 실제 현재까지의 총 관객수 14,395,326, $\hat{m}$은 각 모형으로 추정한 총 관객수이다. 4가지 모형의 상대오차를 비교해보았을 때 n=28, Bass모형일 때가 제일 작으므로 예측 모형으로 Bass 모형을 선택하였다.

n	Bass	Logistic	Gumbel	Exponential
7	NA	-69.17581	-56.82009	-129.44485
14	-35.894488	-47.34650	-39.69862	772.41995
28	**-8.56222**	-17.44321	-12.75310	45.26231

3) Q-Q Plot을 이용한 최적 모형의 적절성 검토

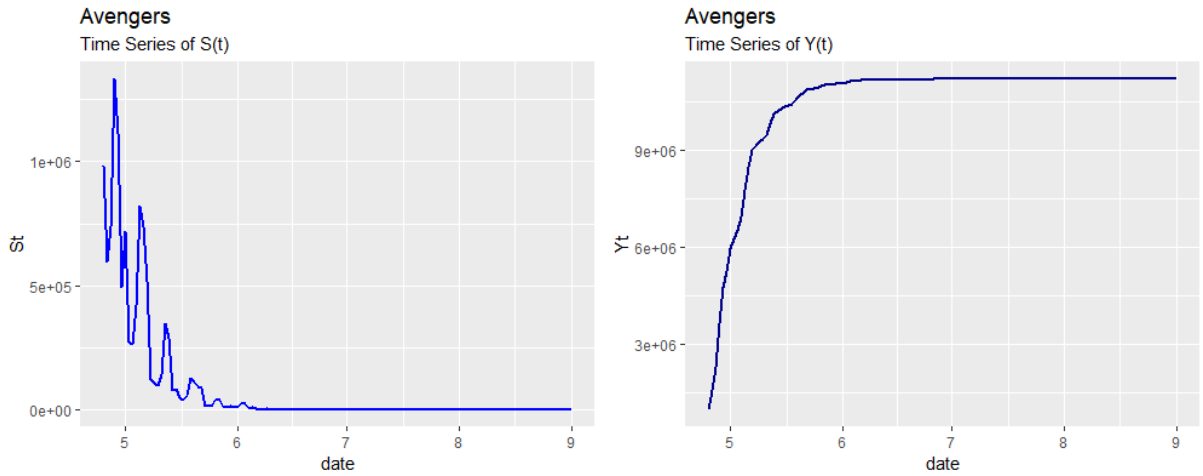



```
##
## Call:
## lm(formula = t ~ 0 + q_bass, data = qqtable1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18.6266   0.5284   2.3618   5.2428   7.1877
##
## Coefficients:
##        Estimate Std. Error t value Pr(>|t|)
## q_bass 0.957986   0.009946   96.32   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.479 on 93 degrees of freedom
## Multiple R-squared:  0.9901, Adjusted R-squared:  0.99
## F-statistic:  9277 on 1 and 93 DF,  p-value: < 2.2e-16
```

➔ Bass 모형에 대한 Q-Q plot을 그려 보았을 때 plot이 직선에 가까워 보이므로 해당 모형이 적절하다고 할 수 있다. 관람시점 t와 bass 분포의 quantile의 선형 회귀모형에 대한 결정계수 값은 99.01%로 매우 높다.

[영화 2] 어벤져스 : 인피니티 워 (2018.04.25 개봉)

1) 개봉 후 일별 관객수 S(t)와 누적 관객수 Y(t)의 시계열 도표 (~2018.09.01)



➔ 어벤져스의 시계열 도표에서 개봉 직후부터 초기의 관객수 St가 매우 높고, 이후 빠른 속도로 줄어드는 것을 확인할 수 있다. 신과 함께와 동일하게 휴일에 대한 보정이 필요해 보인다.

2) 확산 모형을 이용한 최적 예측 모형 탐색

모형을 적합하기에 앞서 공휴일과 주말(토일)의 관객수를 2일로 나누어 계산하도록 하여 휴일효과를 보정하였다. 따라서 보정 전 총 117일의 데이터에 대해 155개 관측치가 있는 수정된 데이터 셋을 생성하고 시점 t를 새롭게 정의하였다. 이 데이터의 처음 n=7, n=14, n=28의 관측치를 사용하여 4가지 OLS 모형을 적합시켜 총 관객수 m을 추정하였다.

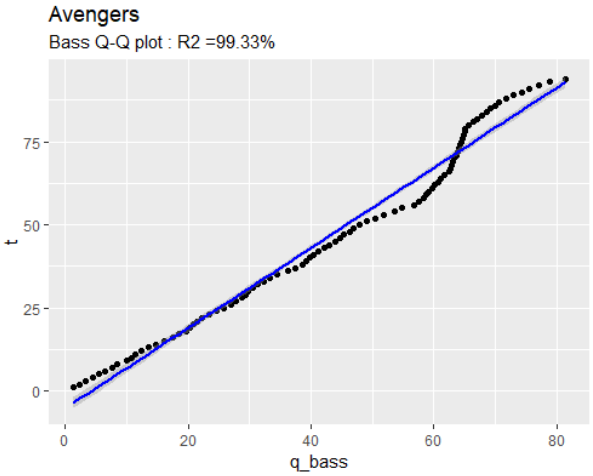
```
##         date      St      Yt  t
## 1  2018-04-25 980042.0  980042  1
## 2  2018-04-26 595668.0 1575710  2
## 3  2018-04-27 739908.0 2315618  3
## 4  2018-04-28 666653.5 2982272  4
## 5  2018-04-28 666653.5 3648925  5
## 6  2018-04-29 557473.0 4206398  6
## 7  2018-04-29 557473.0 4763871  7
## 8  2018-04-30 495787.0 5259658  8
## 9  2018-05-01 717324.0 5976982  9
## 10 2018-05-02 276967.0 6253949 10
```

어벤져스의 경우도 역시 개봉 이후 시간이 많이 지나 현재 박스오피스에서 더 이상 상영하지 않으므로, 상대오차를 계산하기 위해 실제 총 관객수 m에 현재까지 총 누적 관객수로 가정해도 무리가 없다고 판단하고 수정된 데이터에서 Yt의 가장 마지막 값인 **11,211,221**(명)을 사용하였다.

n=7, 14, 28의 Bass, Logistic, Gumbel, Exponential 모형에 대해 총 12개의 모형에 대해 상대오차를 계산한 결과는 위의 표와 같다. n=28인 4주차의 관객수를 이용해 Bass 모형을 적합했을 때의 상대오차 값이 제일 작아 해당 모형을 선택하였다. 해당 모형에서의 모수들의 추정치는 m=10,921,638, p=0.06439, q=0.04973이다. 어벤져스의 경우 추정 총 관객수가 실제 총 관객수 11,211,221명보다 조금 작게 추정되었고 혁신계수 p는 모방계수 q보다 조금 더 크게 추정되었다.

n	Bass	Logistic	Gumbel	Exponential
7	-42.957482	-53.948369	-45.564034	96.642336
14	-8.455311	-28.910562	-23.231043	23.971272
28	**-2.582971**	-9.831206	-7.782529	6.448084

3) Q-Q Plot을 이용한 최적 모형의 적절성 검토



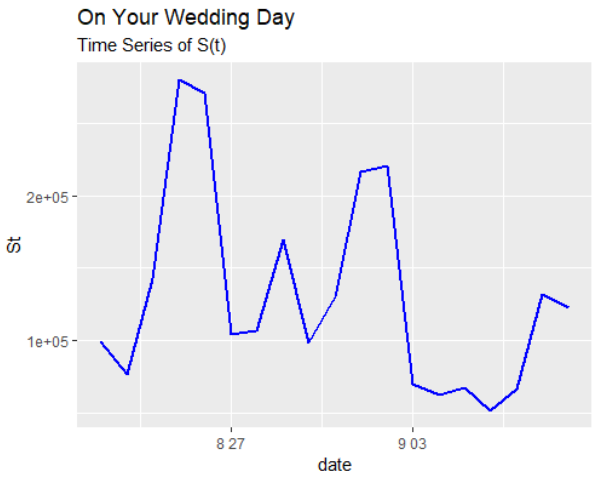

```
##
## Call:
## lm(formula = t ~ 0 + q_bass, data = qqtable2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.1662 -3.8537 -2.9530  0.1001  8.4275
##
## Coefficients:
##        Estimate Std. Error t value Pr(>|t|)
## q_bass 1.111973   0.009493   117.1   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.513 on 93 degrees of freedom
## Multiple R-squared:  0.9933, Adjusted R-squared:  0.9932
## F-statistic: 1.372e+04 on 1 and 93 DF,  p-value: < 2.2e-16
```

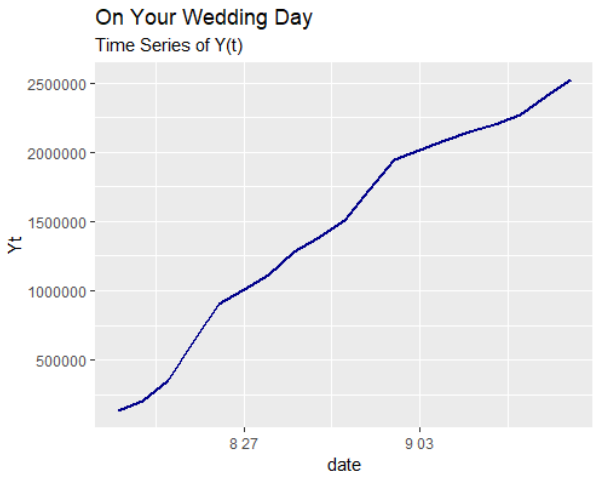
선택된 Bass 모형에 대한 적절성 검토를 위해 실제 총관객수 m=11,211,221을 이용해 관람시점 t를 y축, k-1*ln((1+c*Ur)/(1-Ur)) 을 x축으로 하여 Q-Q plot을 그려보았다. (Bass Quantile을 계산 시에는 실제 m값과 모형에서 추정된 p, q값을 이용해 k=p+q, c=q/p, Ur=Yt/(m+1)의 값들을 각각 계산함)

관측치 중 7월 3일 이후는 스크린 수가 한자리 수이며, 관객수의 관측 날짜 역시 하루 단위가 아니므로 실제 상영 기간에 해당하지 않는다고 판단하여 제외하였다. 그래프를 확인해볼 때 절편을 0으로 한 회귀모형의 결정계수는 99.33%으로 매우 높다. 회귀계수(sigma) 역시 1.11정도로 1에 가까우므로 관람시간 t의 분포가 Bass 분포와 거의 동일하다고 할 수 있다.

[영화 3] 너의 결혼식 (2018.08.22 개봉 ~ 상영 중)

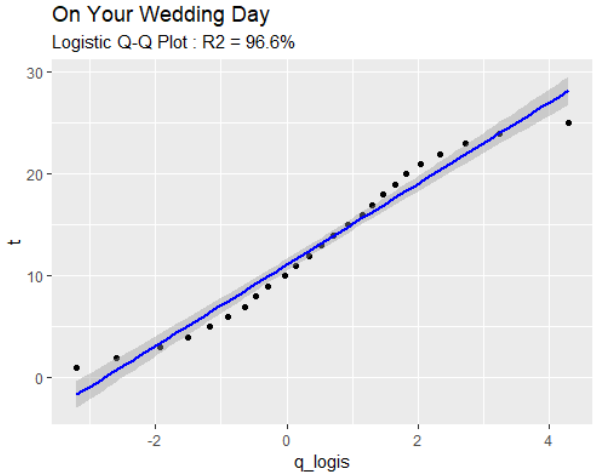
4) 2018년 9월 9일 현재까지 관객수의 시계열 도표와 위의 모형을 이용한 총관객수 추정





<너의 결혼식>은 신과 함께, 어벤져스와는 다르게 현재 상영 중인 영화이므로 S(t)가 급격하게 감소하는 구간이나 Y(t)가 일정해지는 구간을 볼 수 없다. 역시나 휴일효과 보정을 한 후 OLS로 실제 총 관객수를 예측하고자 한다. 9월 9일 데이터가 마지막이므로 9월 9일까지의 누적 관객수를 실제 총 관객수라고 생각하였을 때 구한 상대오차는 다음과 같다. Logistic 모형일 때 상대 오차 값이 가장 작으므로 예측 모형으로 해당 Logistic 모형으로 선택하여 예측된 총관객수 $\hat{m}$을 확인해 보았을 때 **2,525,002**명이라고 생각이 된다.

n	Bass	Gumbel	Logistic	Exponential
7	-53.5108964	-51.750735	-61.748213	-182.52539
14	0.6047819	-15.229131	-24.987338	805.81571
21	-8.45222	7.095895	**1.372972**	76.73734




```
##
## Call:
## lm(formula = t ~ q_logis, data = qqtable3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2011 -1.0186 -0.0019  1.0791  2.6565
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  11.0928     0.2872   38.62   <2e-16 ***
## q_logis       3.9896     0.1561   25.55   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.387 on 23 degrees of freedom
## Multiple R-squared:  0.966,  Adjusted R-squared:  0.9645
## F-statistic: 652.9 on 1 and 23 DF,  p-value: < 2.2e-16
```

또한 Q-Q Plot 역시 거의 직선에 가까운 모양이 나오므로 OLS로 선택된 Logistic 모형이 적절하다고 생각할 수 있다. Logistic quantile 값과 t의 선형 회귀모형에서 계수의 p-value 값이 매우 작고 R^2 역시 96% 정도로 높다.