이론통계학2

Project #1. 신상품 수요예측, 전염병 확산 예측모형

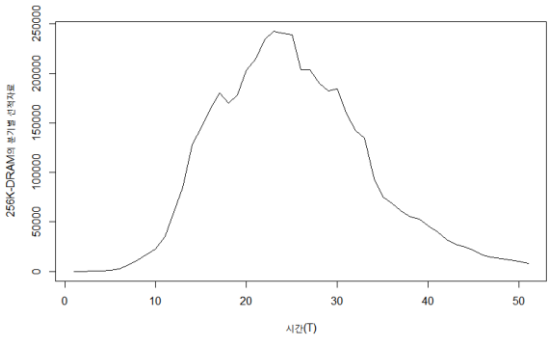
발표일 : 2021.09.15

1조

202STG26	박지윤
202STG27	이수현
212STG04	김이현
212STG12	박윤정

Part A : DRAM 분기별 선적자료(1982-1995)

1. 256K-DRAM 분기별 선적자료에 대한 시계열도표를 그려보시오.



점차 증가하다가 T=23 부근에서 피크값을 가진 후 점차 감소하는 추세를 보인다.

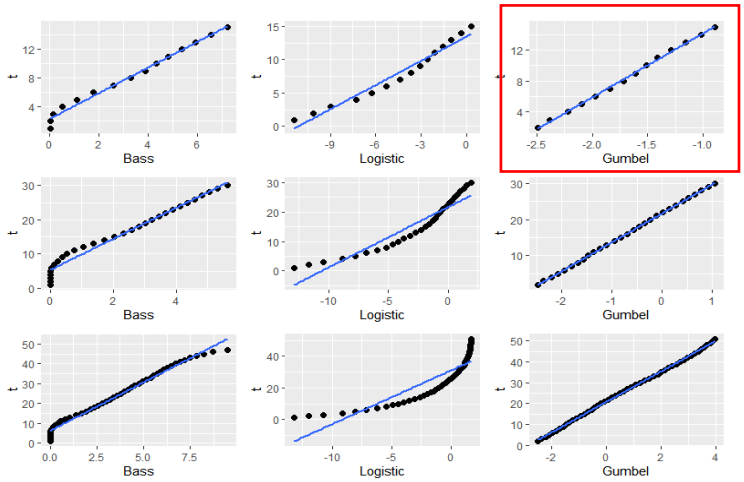
2. 256K-DRAM 의 분기별 선적자료를 이용하여 DRAM 의 총수요(m)을 추정하려한 다. 아래 세 모형을 대하여 최초 n=15,30,51 개의 자료를 학습자료(train data)로 이용하여 각 모 형에 포함된 모수 (p,q,m)들을 OLS 방법으로 추정하고 상대오차값을 구하시오.

n	model	p	q	m	상대 오차
15	bass	0.001	0.69	892463	-81.01
	logistic	0	0.7	881246	-81.25
	gumbel	.	0.15	5890455	25.33
30	bass	0.005	0.23	4147325	-11.76
	logistic	0	0.25	4049601	-2
	gumbel	.	0.13	5007992	6.55
51	bass	0.01	0.19	4621534	-1.70
	logistic	0	0.21	4606018	-2
	gumbel	.	0.14	4740560	0.86

OLS 방법으로 모수를 추정한 결과표이다. 이후 실제총수요는 n=51 때의 총 누적판매량보다 약간 큰 점을 고려해 실제 총수요 m=4700000으로 정한 후 상대오차값을 구해주었다. 이 때 n=51일 때의 Gumbel 모형의 상대오차값이 가장 작고 이 때의 모수는 q=0.14, m=4740560으로 추정된다.

3. MSE 및 Q-Q Plot을 이용한 최적 예측 모형 선택

n	Bass	Logistic	Gumbel
15	10440478	10650682	24716343
30	513619976	707073569	168957327
51	439914408	653499804	147434250



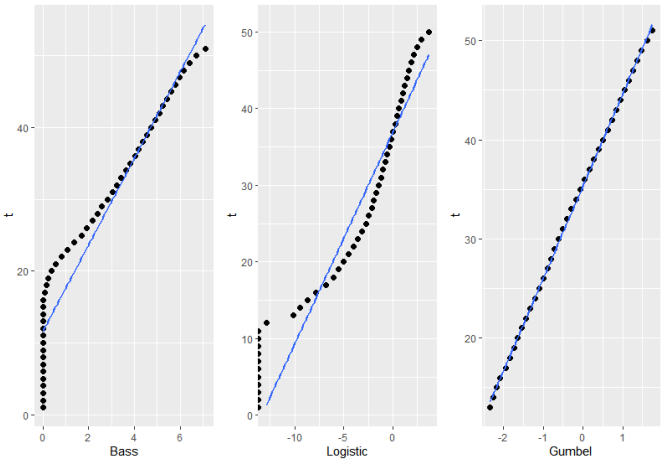
모형 적합 후 최적 모형 판단을 위해 각 모델의 MSE를 구하고 Q-Q plot을 그려보았다.Q-Q plot에서 Bass의 경우 선형관계를 확인할 수 있으나 n=15, 30일 때 왼쪽 끝부분, n=51일 때 양쪽 끝부분이 선형에서 벗어난다. 세 경우 모두 0.97 이상의 높은 R^2값을 보였다. Logistic의 경우 n=15,30일 때는 약한 선형관계를 확인할 수 있고 R^2값도 0.9 이상이었으나 n=51의 경우 선형을 크게 벗어났고 R^2값또한 0.77로 낮았다. Gumbel의 경우 세 경우 모두 강한 선형관계를 보였으며 R^2값도 전부 1에 가까운 값을 가졌다. MSE는 n=15 일 때의 Bass 가장 작았다. 하지만 해당 모델의 경우 Q-Q plot에서 왼쪽 끝이 잘 적합 되지 않았기 때문에 모든 경우가 잘 적합 된 Gumbel 모형을 택했고, 이 중 MSE가 가장 작은 n=15일 때의 모형을 최적 모형으로 판단했다. 이 때의 상대오차값은 25.33이다.

4. 1M-DRAM 자료를 이용한 m 추정

DRAM Qu	Quantity	T(시점)	1MD
82	1	1	0
	2	2	0
	3	3	0
	4	4	0
83	1	5	0
	2	6	0
	3	7	0
	4	8	0
84	1	9	0
	2	10	0
	3	11	0

t=1 부터 t=11 까지의 값이 누락된 자료인 1M-DRAM 자료를 이용하여 위와 마찬가지로 OLS 를 통해 모수를 추정한 후 Q-Q plot 을 그려보았다.

<Q-Q plot>



	Bass	Logistic	Gumbel
m	4539646	5236765	5093646

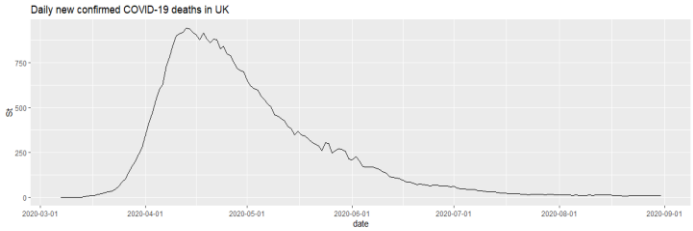
1M-DRAM 의 자료의 경우 t=11 까지의 값이 0 이므로 Gumbel 모형은 t=12 부터의 자료를 사용했다. Bass 의 경우 왼쪽 끝부분과 오른쪽 끝부분에서 선형관계가 약간 벗어나나 대체적으로 선형을 유지한다. R^2은 0.93 이다. Logistic 의 경우 선형관계가 강하지 않으며 R^2값은 0.91 로 세 모델 중 가장 낮았다. Gumbel 의 경우는 강한 선형관계를 육안으로도 확인 가능하며 R^2값도 1 에 가까워 적절한 모형임을 판단했다. 세 경우 모두 n=41 때의 누적판매량인 m=4176735 보다 큰 값으로 m 이 추정되었으며, 이는 자료가 censored 된 자료임을 고려했을 때 적절히 추정된 값이라고 판단할 수 있다.

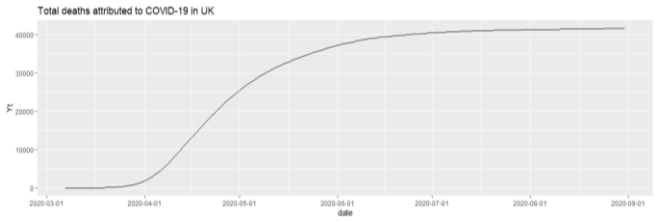
PartB-1: Covid-19 사망자 예측

1. 2020.3.7 ~ 2020.8.31 영국의 일별 Covid-19 사망자 수

(rolling 7-day average)

a) 위 사이트에서 다운받은 엑셀자료를 이용하여 최초 사망자가 발생한 날부터 8 월 31 일까지 일별 신규 Covid-19 사망자수 자료 및 누적 사망자수 자료에 대한 시계열도표를 그려보시오.





2020 년 3 월 7 일 영국에서 코로나 19 로 인한 사망자가 발생한 후 2020 년 4 월 13 일까지 일일 신규 사망자가 급격하게 증가하다가 이후 천천히 감소하며 오른쪽 꼬리가 긴 분포를 보이고 있다. 또한 누적 사망자 수도 시간이 지남에 따라 수렴하는 양상을 보인다.

b) 최초 사망자가 발생한 날 부터 초기 n 일 까지 일별 Covid-19 사망자수 자료를 학습자료(train data)로 이용하여 장차 영국 내 총 감염자수(m)을 추정하려한다.

가) n= 20,30,50 일 경우 세 가지 모형을 이용한 해당 모수 (p,q,m)들을 각각 추정하고 추정된 m 값과 최신 m 값(예; 8 월 31 일까지 누적 사망자수) 과 비교한 상대 오차값을 구하고 그 의미를 설명하시오.

나) 이들 모형 중 최적 모형을 선택하고 그 선택 근거를 설명하시오.

다) 위에서 선택된 모형에 대해 여러 추정 방법 (예: OLS, Q-Q plot 방법, NLSE, MLE, Bayesian 추정 등)을 이용한 m 추정값의 정확도를 서로 비교해 보고각 방법의 장단점을 기술하시오.

1. OLS 추정법

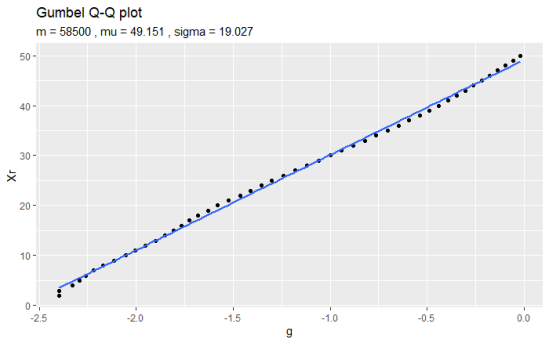
각각 n=20, 30, 50 일 때 Bass, Logistic, Gumbel model 을 가정하여 OLS 추정법을 통해 모수(m,p,q)를 추정하였다. 아래 표는 모수 추정 결과와 영국의 2020 년 8 월 31 일의 누적 사망자 수(m=41569)와 비교한 상대 오차를 구한 결과이다. Covid-19 사망자 수는 혁신 계수(p)보다 모방 계수(q)가 더 크다. Covid-19 감염은 접촉에 의해 발생하기 때문이다. 영국의 신규 코로나 사망자 수가 2020 년 4 월 중순에 정점을 찍는다. n 이 작을 경우 학습자료에 정점을 포함하지 못하므로 m 이 작게 추정된다. 또한 신규 코로나 사망자 수는 정점 전후의 증가세(감소세)가 다르다. 이러한 분포를 normal 분포의 모양을 가지는 Bass 와 Logistic model 은 고려하지 못하므로 해당 데이터에는 적절하지 않다. 이로 인해 상대 오차가 Gumbel 에 비해 크다. 따라서 꼬리가 긴 분포를 가지며 정점을 학습자료에 포함하는 n=50 이고 Gumbel model 을 이용한 모형을 최적으로 선택한다.

표 1. OLS 추정 결과

n	model	p	q	m	상대 오차
20	bass	0	0.284	-47098.6	-213.302
	logistic	0	0.312	5014.32	-87.937
	gumbel	•	0.024	49433347	118818.8
30	bass	0	0.276	9694.136	-76.679
	logistic	0	0.281	9354.834	-77.496
	gumbel	•	0.052	122390.9	194.428
50	bass	0.002	0.157	25462.15	-38.747
	logistic	0	0.166	25013.73	-39.826
	gumbel	•	0.078	32854.69	-20.963

2. Q-Q plot 추정법

OLS 추정법에서 선택한 모형인 n=50, Gumbel 모형을 이용하여 Q-Q plot 추정법으로 모수를 추정하였다. OLS 추정법에서 m=32854.69 으로 추정된 것을 고려하여 m=30000, 30500, 31000, 31500, …, 70000 일 때 r 번째 사망자의 사망시간을 X(r)을 종속변수로 G-1Ur= G-1r/(m+1)을 독립변수로 하여 선형회귀모형을 적합시켰다. 그 결과 R2 를 극대화시키는 m 값은 58500 이다. 그 때의 μ=49.151, σ=19.027 이다.



3. NLSE

n=50, Gumbel 모형을 기반으로 OLS 추정치와 Q-Q plot 추정치를 초기값으로 하여 NLSE 를 구하였다. 모형 1 로 추정한 결과는 m=32854.69, q=0.078 이고, 모형 2 의 경우 m=58486.77, μ=49.147, σ=19.025 이다.

- 모형 1: $St = aYt - 1 + bYt - 1 \cdot lnYt - 1 + et, \quad a = q \cdot ln\,(m)\,, b = -q$
- 모형 2: $Xr = \mu + \sigma(-\ln -\ln Ur \;\; + er, \quad Ur = r/(m+1)$

4. MLE

n=50, Gumbel 모형의 log-likelihood 를 구하여 optim 함수를 통해 이를 maximize 하는 모수 MLE 를 추정하였다. 그 결과 m=58459.011, μ=49.945, σ=20.007 이다.

5. 비교

모형 1 을 기반으로 추정한 OLS 와 NLSE 는 상대 오차는 약 -21%이고 모형 2 를 기반으로 추정한 Q-Q plot, NLSE, MLE 상대 오차가 약 41% 이다. 전자의 경우 전체 사망자 수를 적게 추정하였고 후자의 경우 더 많게 추정하였다. 절대값은 전자가 더 작으므로 정확도는 더 높았다. 또한 같은 모형을 이용할 때 추정 방법 간의 차이는 크게 없다.

라) 이태리의 일별 Covid-19 사망자수 자료를 이용하여 최초 n=20,30,50 일의 학습자료를 이용하 여 각 경우 최적 예측모형을 찾아보고 서로 다른 추정 방법의 정확도를 비교해 보시오.

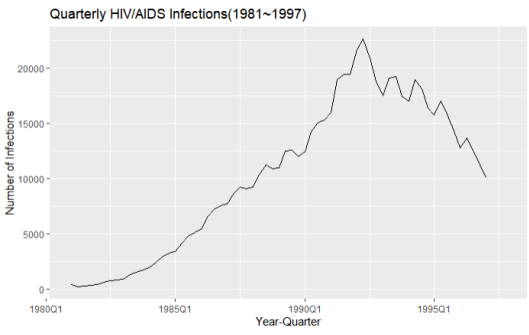
이탈리아의 Covid-19 사망자 수 분포도 영국과 동일하다. Covid-19 일일 사망자 수는 급격하게 증가하다가 2020 년 4 월 중순에 정점을 찍고 비교적 완만하게 감소한다. 따라서 오른쪽 꼬리가 긴 분포로 gumbel 모형이 적절할 것이다. n=20, 30, 50 일 때 Bass, Logistic, Gumbel 모형을 이용하여 OLS 추정법으로 모수를 추정하였다. 2020 년 8 월 31 일의 이탈리아의 누적 사망자 수(m=35461.29)과 비교하였을 때 상대 오차는 n 이 커질수록 절대값이 작아졌고 gumbel 모형이 가장 작았다. 영국과 이탈리아의 코로나 사망자 수의 분포는 거의 동일한 양상을 보이기에 이러한 결과가 나왔을 것이다.

N=50 까지를 학습데이터로, gumbel 모형을 이용하여 Q-Q plot, NLSE, MLE 를 구하였다. 그 결과 m 은 각각 48000 (Q-Q plot), 26636.6 (NLSE-모형 1), 48037.7 (NLSE-모형 2), 47752.096 (MLE)이다. 상대 오차는 OLS, NLSE-모형 1 의 경우 약 -25%, Q-Q plot, NLSE-모형 2, MLE 의 경우 약 35%이다. 영국과 동일하게 전자의 경우 총 사망자 수를 적게 추정하나 상대 오차는 적고, 후자의 경우 상대 오차는 크나 총 사망자 수를 많게 추정한다.

Part B-2 : 미국과 한국의 HIV/AIDS 확산 예측

1) 미국 분기별 HIV/AIDS-감염자자료 (1981-1997):

a) 분기별 HIV/AIDS 감염자 자료에 대한 시계열 도표를 그려보시오.



미국 HIV/AIDS 감염자 수는 1981 년도부터 서서히 증가하다가 1991 년도에 약 23,000 명을 기록하며 그래프는 정점을 찍는다. 이후 1997 년도까지 분기별로 증감을 반복하지만 전체적으로 감소 추세를 보인다.

b) OLS 추정

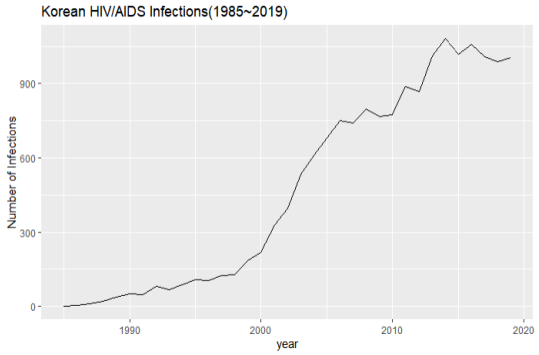
n 이 20, 40, 65 일 경우 Bass, Logistic, Gumbel 모형에 대한 모수 (m, p, q)와 2019 년도 실제 누적 감염자 수(1,189,700 명)와 비교한 상대오차값을 구한 결과는 다음과 같다.

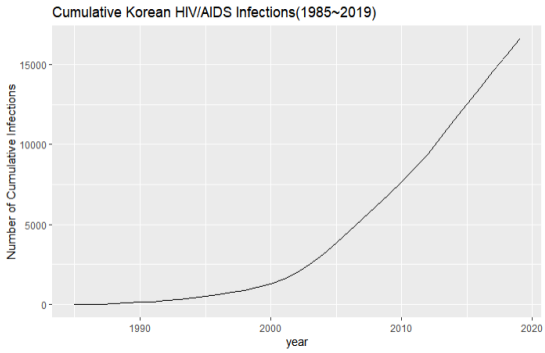
n	20			40			65		
model	Bass	Logistic	Gumbel	Bass	Logistic	Gumbel	Bass	Logistic	Gumbel
p	0.002	0	-	0.002	0	-	0.002	0	-
q	0.214	0.246	0.05	0.12	0.137	0.046	0.097	0.103	0.048
m	116136.6	88571.83	784234.6	450452.1	403335.9	886443.9	78640.2	773274.8	10000860
상대 오차	-90.238	-92.555	-34.081	-62.137	-66.098	-25.49	-33.871	-35.003	-15.873

모든 경우에서 q 값(모방 계수)이 p 값(혁신 계수)보다 크다. 이는 감염병인 HIV/AIDS 가 접촉을 통해 발생하기 때문이다. n 이 20 일 때보다 65 일 때, Gumbel 모형으로 추정하였을 때 상대오차가 작은 것을 확인할 수 있다. Gumbel 분포는 다른 분포에 비해 꼬리가 긴 편이기 때문에 이러한 감염병 분포에 적합하다고 볼 수 있다. 그러므로 상대오차가 가장 작고 분포의 의미를 내포하는 n=65 일 때의 Gumbel 분포를 최적의 모형으로 선택한다.

2) 국내 연도별 HIV/AIDS 감염 현황 자료 (1985-2020) – 총 감염자 수 추정

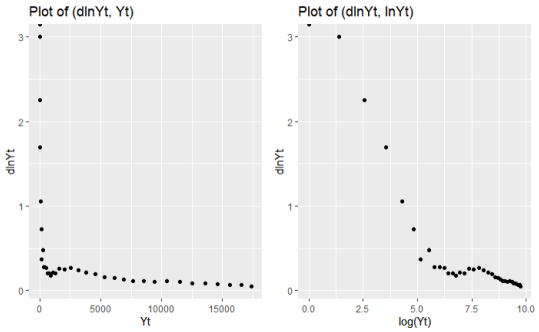
a) 국내 HIV/AIDS 감염자(t<=2019) 시계열 도표





국내 HIV/AIDS 감염자 수는 1985 년도부터 서서히 증가하다가 2000 년도를 기준으로 급증하였다. 하지만 2015 년도부터 현재까지 미미한 감소세를 보이고 있다. 누적 감염자 수의 경우 2000 년도를 기점으로 급증하였으며 꾸준한 증가세를 보인다.

b) OLS 추정

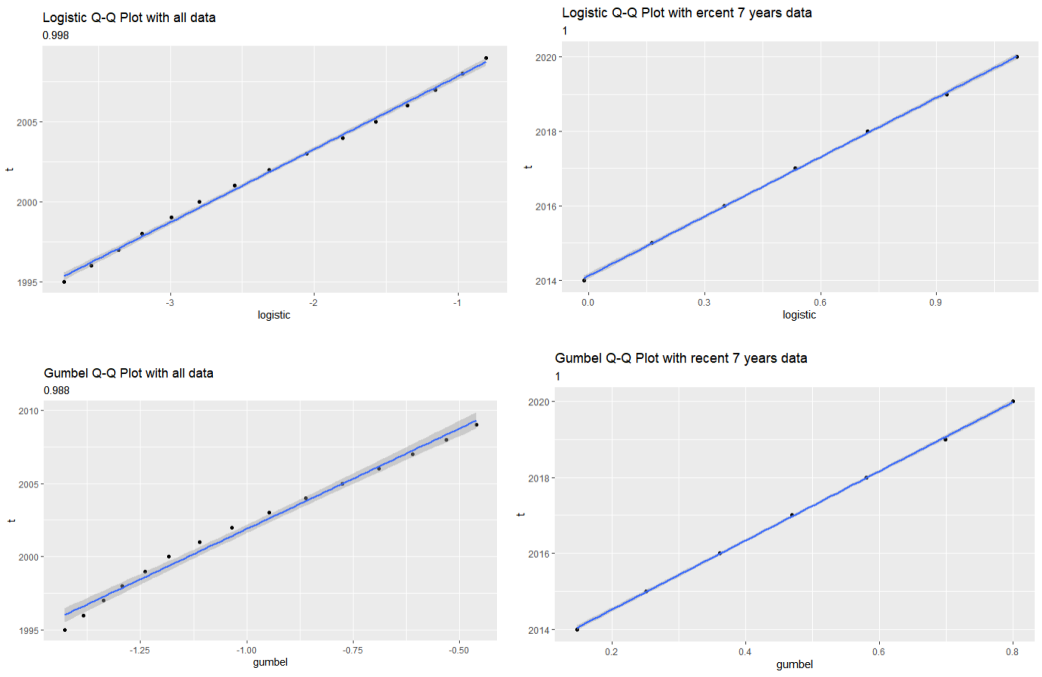


왼쪽 그래프는 dlnYt 와 Yt, lnYt 의 산점도를 그린 결과이다. Y(t)를 로그 변환했을 때 더 선형임을 볼 수 있다.

	방법 1(전체 데이터 사용)		방법 2(최근 7 년 데이터 사용)	
model	Logistic	Gumbel	Logistic	Gumbel
q	0.193	0.081	0.183	0.107
m	22264.06	33501.29	23144.83	27264.78

c) Q-Q plot 추정

두가지 방법과 두 모형을 이용한 결과는 다음과 같다.

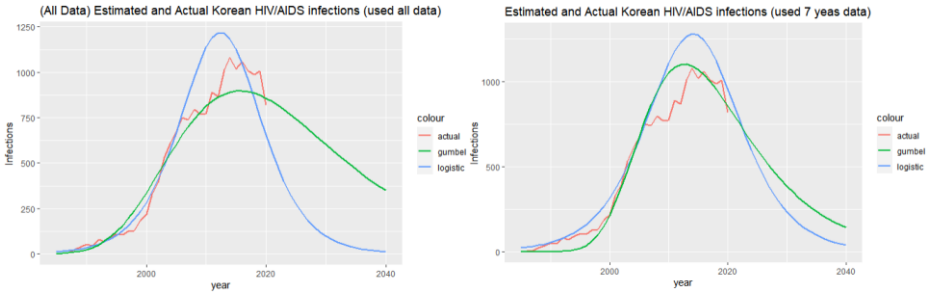


	방법 1(전체 데이터 사용)		방법 2(최근 7 년 데이터 사용)	
model	**Logistic**	**Gumbel**	**Logistic**	**Gumbel**
m	22264.060	33501.290	23144.830	27264.780
mu	2012.418	2015.627	2014.114	2012.697
sigma	4.566	13.735	5.322	9.102

4 개의 Q-Q plot 중 R2 값이 가장 높은 최근 7 년간의 자료를 사용한 Logistic 모형과 Gumbel 모형을 선택한다. 그에 따라 추정 모수는 Logistic 모형의 경우 m=23144.830, mu=2014.114 이며 Gumbel 모형의 경우 sigma=5.322, m=27264.78, mu=2012.697, sigma=9.102 이다.

d) 미래 HI V 감염자 수 예측

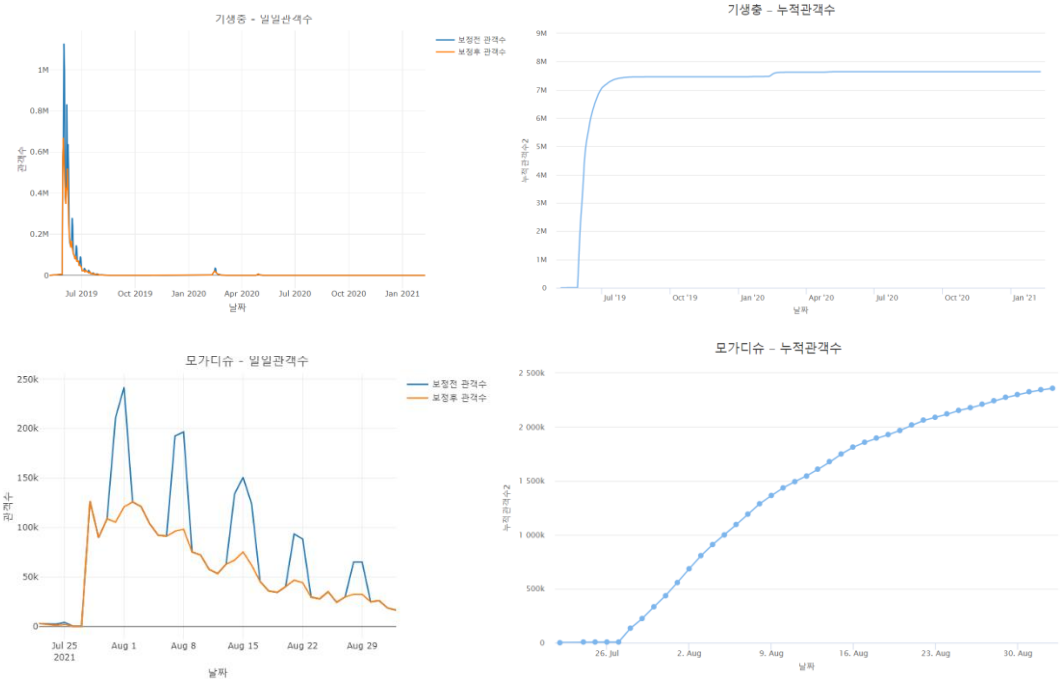
전체 데이터로 사용했을 때와 최근 7 년간의 데이터를 사용했을 때의 모수값을 통해 1985~2040 년의 국내 HIV/AIDS 감염자 수를 예측하였다. 이를 실제 관측값(~2020 년)과 함께 비교하였으며 결과는 다음 그래프와 같다.



전체 데이터를 사용 후 모수를 추정한 결과, Gumbel 과 Logistic 모형 모두 초기값은 예측을 잘했으나 정점 부근에서부터 실제값과의 차이가 커진다. 반면, 최근 7 년간 데이터를 사용 후 모수를 추정한 결과, 초기값은 예측이 상대적으로 떨어지지만 정점 부근에서 Gumbel 모형의 예측력이 높아지는 것을 볼 수 있다. 전체 데이터를 사용했을 때보다 비교적 직선 형태의 상승세를 보이던 최근 7 년간의 데이터를 사용했을 때, Gumbel 모형의 예측력이 높다고 할 수 있다.

Part C : 영화 흥행 예측

1. 개봉 후 현재까지 일별 관객수 및 누적관객수 시계열 도표를 그리시오.



영화 관객수의 경우, 주말과 평일의 관객수 차이가 매우 크기 때문에, 예측의 편의를 위하여 주말을 이틀에 할당하여 실제 주말 관객수의 1/2 을 하루 관객수로 놓고 추정하였다. 주말 보정을 한 결과, 고점이 많이 사라진 것을 확인할 수 있다.

2) 아래 4 가지 확산모형과 개봉 후 1 주, 2 주 및 4 주 간 흥행자료를 이용하여 총관객수(m)을 추정한 후 이를 실제총관객수 m 값과 비교한 상대 오차 값을 구하여 최적 예측 모형을 찾아보시오.

- <기생충> 7 일간의 데이터를 이용한 4 가지 모형의 모수 추정 (m = 10,313,163)

모형	$\hat{m}$	$\hat{p}$	$\hat{q}$	상대오차
Bass	5,243,704	0.114	0.149	-49.155
logistic	3,881,936	0.000	0.673	-62.359
Gumbel	4,238,285	.	0.416	-58.904
Exponential	10,601,146	0.060	-	2.792

- <기생충> 14 일간의 데이터를 이용한 4 가지 모형의 모수 추정 (m = 10,313,163)

모형	$\hat{m}$	$\hat{p}$	$\hat{q}$	상대오차
Bass	-	-	-	-
logistic	7,045,548	0.000	0.307	-31.684
Gumbel	7,571,702	.	0.197	-26.582
Exponential	13,898,894	0.044	0.000	34.768

- <기생충> 28 일간의 데이터를 이용한 4 가지 모형의 모수 추정 (m = 10,313,163)

모형	$\hat{m}$	$\hat{p}$	$\hat{q}$	상대오차
Bass	9,343,507	0.065	0.040	-9.402
logistic	8,490,846	0.000	0.225	-17.67
Gumbel	8,696,525		0.159	-15.675
Exponential	10,282,634	0.065		-0.296

영화 <기생충> 의 경우, Exponential 모델이 상대 오차 -0.296 로, 실제 흥행 수요에 가장 근접하게 예측하였다.

- < 모가디슈 > 7 일간의 데이터를 이용한 4 가지 모형의 모수 추정 (m = 3,144,878)

모형	$\hat{m}$	$\hat{p}$	$\hat{q}$	상대오차
Bass	720,962	0.053	0.512	-77.075
logistic	610,038	0.000	0.792	-80.602
Gumbel	735,290		0.428	-76.62
Exponential	음수로추정	-	-	-

- <모가디슈> 14 일간의 데이터를 이용한 4 가지 모형의 모수 추정 (m = 3,144,878)

모형	$\hat{m}$	p̂	$\hat{q}$	상대오차
Bass	1,599,528	0.029	0.254	-49.139
logistic	1,458,915	0.000	0.375	-53.61
Gumbel	1,702,159	.	0.206	-45.875
Exponential	음수로 추정		0.000	-

- < 모가디슈 > 28 일간의 데이터를 이용한 4 가지 모형의 모수 추정 (m = 3,144,878)

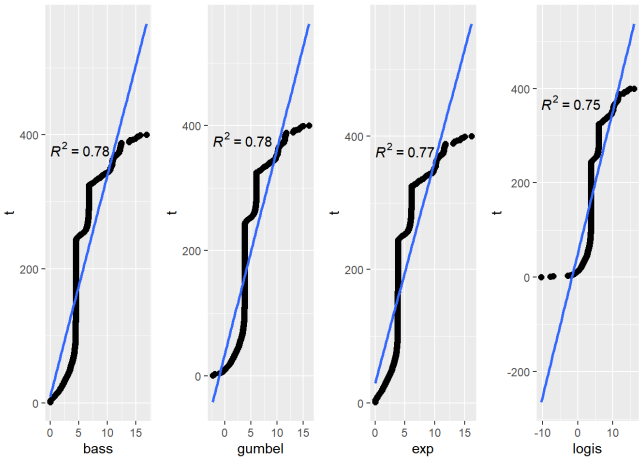
모형	$\hat{m}$	p̂	$\hat{q}$	상대오차
Bass	2,705,494	0.025	0.104	-13.971
logistic	2,469,601	0.000	0.189	-21.472
Gumbel	2,639,939	.	0.122	-16.055
Exponential	12,881,618	0.007	0.000	309.606

영화 <모가디슈> 의 경우, BASS 모델이 상대 오차 -13.971 로, 실제 흥행 수요에 가장 근접하게 예측하였다.

기생충의 경우에는 개봉 전부터 많은 이슈가 있던 영화로 초반에도 큰 흥행을 이끌었다. 따라서 exponential 모형이 선택된 것이 적절하다고 볼 수 있다. 영화 모가디슈의 경우에는, 혁신 계수보다 모방 계수가 전체적으로 큰 것을 확인할 수 있다. 따라서 모가디슈의 경우 입소문 효과가 관객의 수에 영향을 주었다고 볼 수 있다.

3) 실제 총 관객수 m 을 이용하여 해당모형의 Q-Q plot 을 그려보고 해당모형이 적절한지 검토하시오.

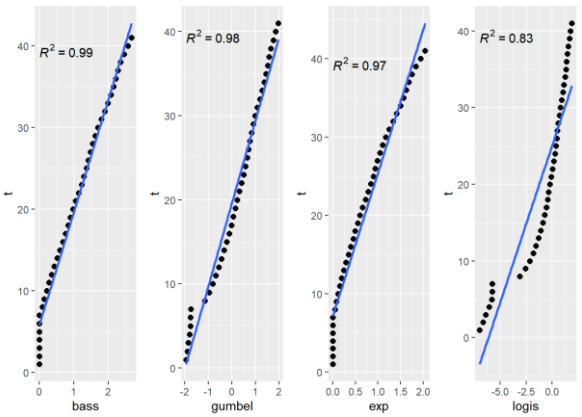
- <기생충> Q-Q plot



직선보다는 여러 개의 곡선을 합쳐 놓은 듯한 모양이 나오는데, 이는 주말 보정으로도 해결하지 못한 고점 값 때문인 것으로 보인다. 따라서 추가적인 주말 보정이 필요한 것으로 확인된다.

- <모가디슈> Q-Q plot

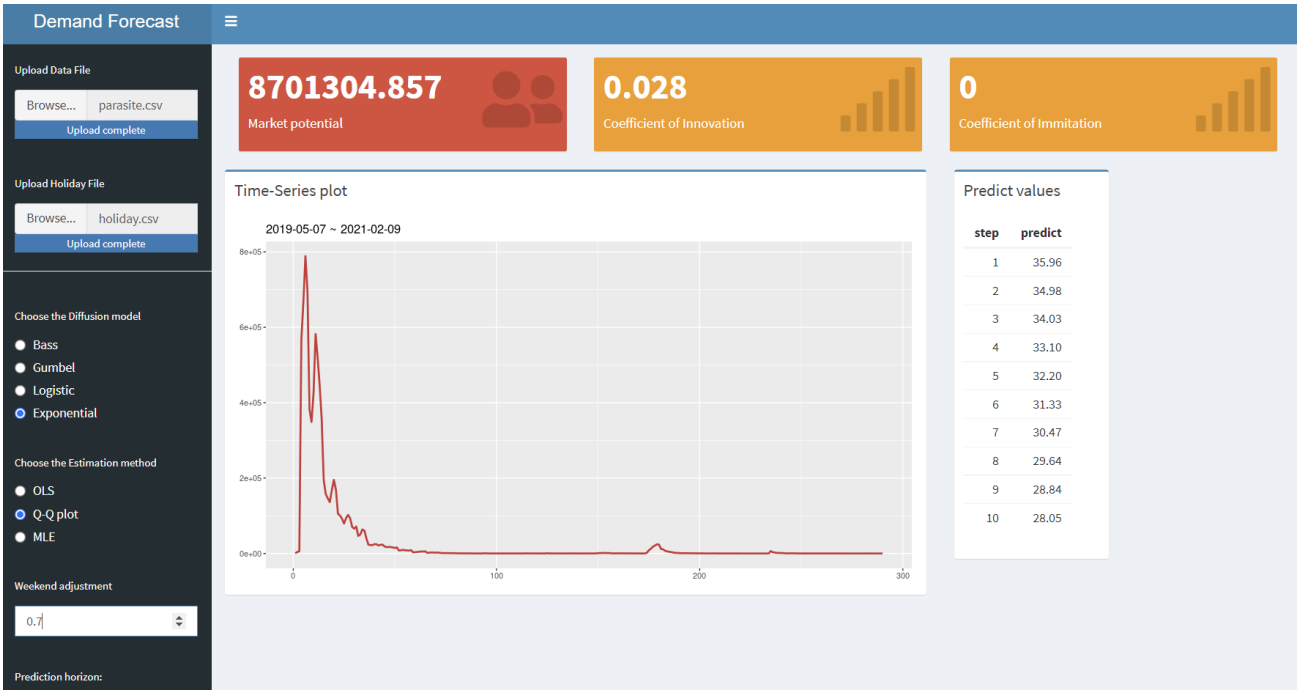
영화 <모가디슈> 의 경우, Q-Q plot 을 살펴보면, Bass 모형이 직선으로 가장 잘 적합된 것을 알 수 있는데, 이는 앞의 모수 추정 방식에 따라서 상대 오차를 비교해본 결과와도 일치한다. 따라서 Bass 모형이 <모가디슈> 관객 수 예측에 적합한 것을 알 수 있다.



Part D : R-shiny 수요예측 App

https://jiyoon-ing.shinyapps.io/project1/

예) 기생충 데이터, 주말 보정 0.7 으로 한 후, Exponential 모델의 Q-Q plot 추정치



예) 국내 HIV 감염자 데이터, Logistic 모형의 MLE 추정치