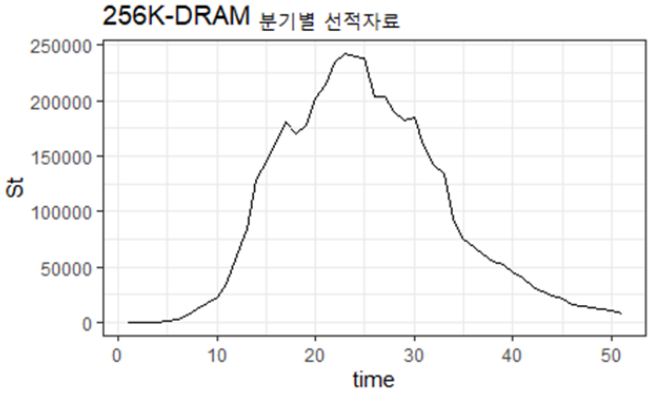
이론통계학2 - project #1 신상품 수요 예측 / 전염병 확산 모형

강아미, 고유정, 윤보인, 이혜린, 홍지원

<Part A> DRAM 분기별 선적자료(1982-1995)

1. 256K-DRAM 분기별 선적자료에 대한 시계열 도표를 그려보시오.



1982 년부터 1995 년까지의 DRAM 의 분기별 판매 수량이다. 1988 년도에 가장 수요가 많았으며 peak 이후로는 계속 감소하는 패턴을 보이고 있다.

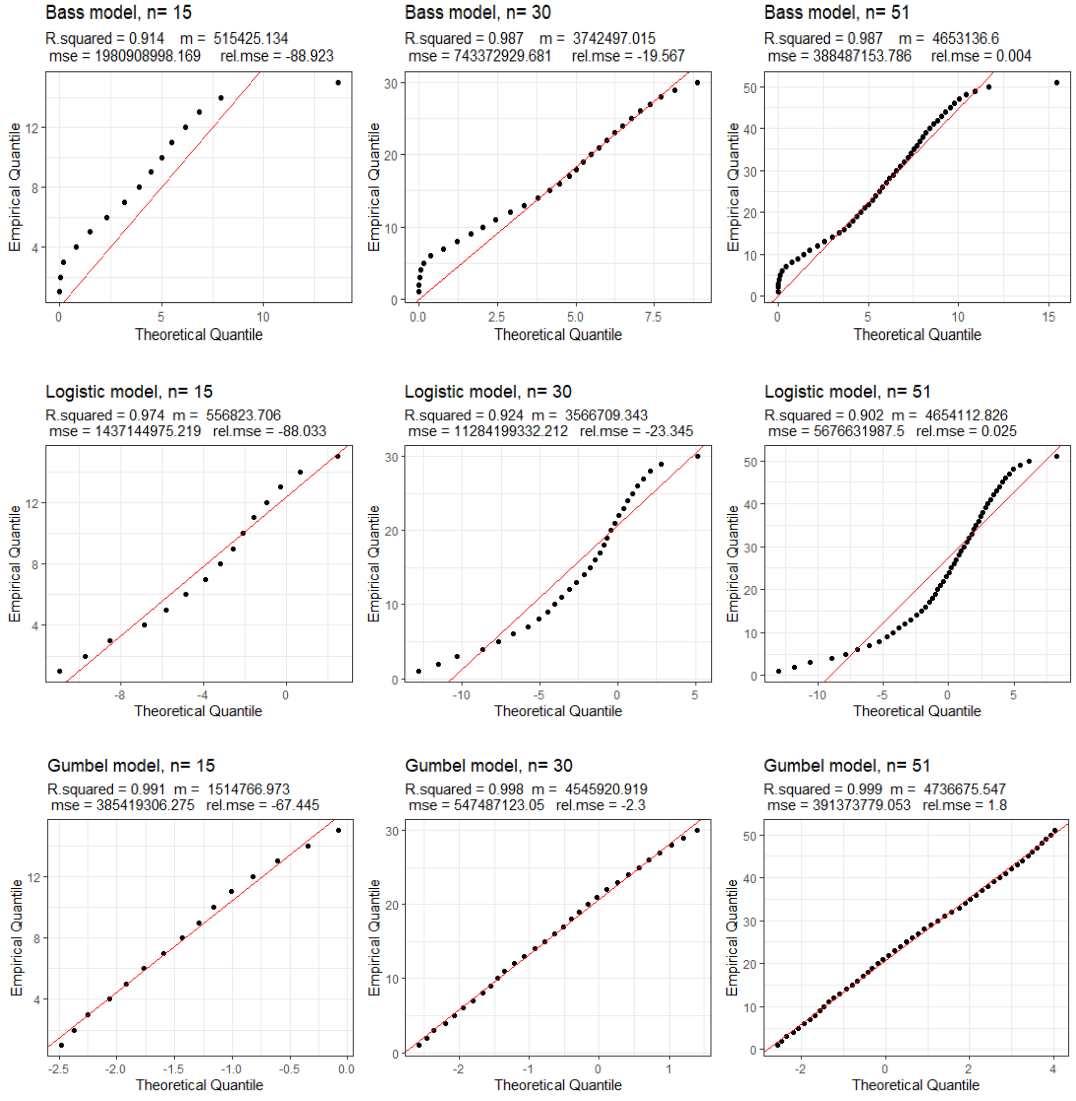
2. 256K-DRAM 의 분기별 선적자료 S_t,t=1,2,…,n 를 이용하여 DRAM 의 총수요(m)을 추정하려 한다. 아래 세 모형을 이용하여 n=15,30,51 일 경우 각 모형에 포함된 모수(p, q, m)들을 OLS 방법으로 추정하고 상대오차값 100*(m^-m)/m 을 구하시오.

n	Bass	Logistic	Gumbel
	$(\hat{m}, \hat{p}, \hat{q})$	$(\hat{m}, \hat{q})$	$(\hat{m}, \hat{q})$
15	(892463, 0.00067, 0.681) 상대오차 = -80.82	(881245.6, 0.688) 상대오차 = -81.06045	(5890455, 0.1487) 상대오차 = 26.59649
30	(4147325, 0.00497, 0.2297) 상대오차 = -10.866	(4049601, 0.2512) 상대오차 = -12.96677	(5007992, 0.127) 상대오차 = 7.630766
51	(4621533, 0.00583, 0.19418) 상대오차 = -0.675	(4606018, 0.2108) 상대오차 = -1.008375	(4740560, 0.1362) 상대오차 = 1.88317

Bass, Logistic, Gumbel 모형을 이용해 256-DRAM 의 누적 수요량을 예측한 결과는 위와 같다. 상대오차를 계산하기 위해 사용된 m 값은 우리가 가진 자료의 총 수요를 합한 값인 4652937 로 실제 누적판매량은 이것보다 클 것이다.

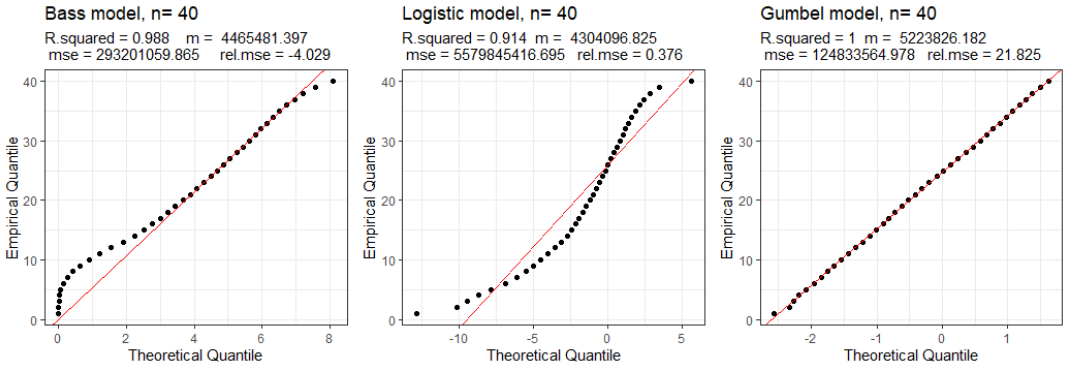
n=15, n=30 일 경우, Gumbel 모형을 사용했을 때 상대오차가 가장 작은 것을 확인할 수 있다. n=51 인 경우에 Logistic 모형이 가장 작은 값이 나왔지만, 전반적으로 세 모형 모두 상대오차 값이 작은 결과가 나왔다. n 의 값의 관계없이 예측을 잘 한 모형은 Gumbel 모형이다.

3. n=15, 30, 51 의 각 경우 MSE, Q-Q plot 등을 이용하여 최적 예측모형을 선택하시오. 또한 선택된 모형을 이용한 총수요 추정치(m^)를 실제총수요(m)와 비교한 상대오차를 구하여 가장 정확한 추정 방법은 무엇인지 설명하시오. (단 실제총수요(m) 은 n=51 일 때의 총 누적판매량 보다 약간 큰 점을 고려할 것)



OLS 방법과 마찬가지로 n 이 클수록 상대오차가 작은 것을 확인할 수 있다. Q-Q plot 을 살펴보면 Logistic 분포는 가장 많은 데이터를 가지고 분석했을 때 R-squared 가 높은 값이 나왔다. Gumbel 분포를 사용했을 때에는 n 값에 관계없이 모두 Q-Q Plot 이 직선 형태를 나타내고 있으며 적절한 분포임을 알 수 있다.

4. 1M-DRAM 전체자료 (n=40)에 대해 위 세 가지 방법으로 예측한 m값을 상호 비교해 보시오. 또 각 모형에 대한 Q-Q plot을 그리고 이들 중 가장 적절한 모형은 무엇인지 검토하시오. (주의: 이 경우 판매자료는 censored 자료여서 실제총수요 m 값은 미지수이고 n=40까지 누적판매량보다 큼.)

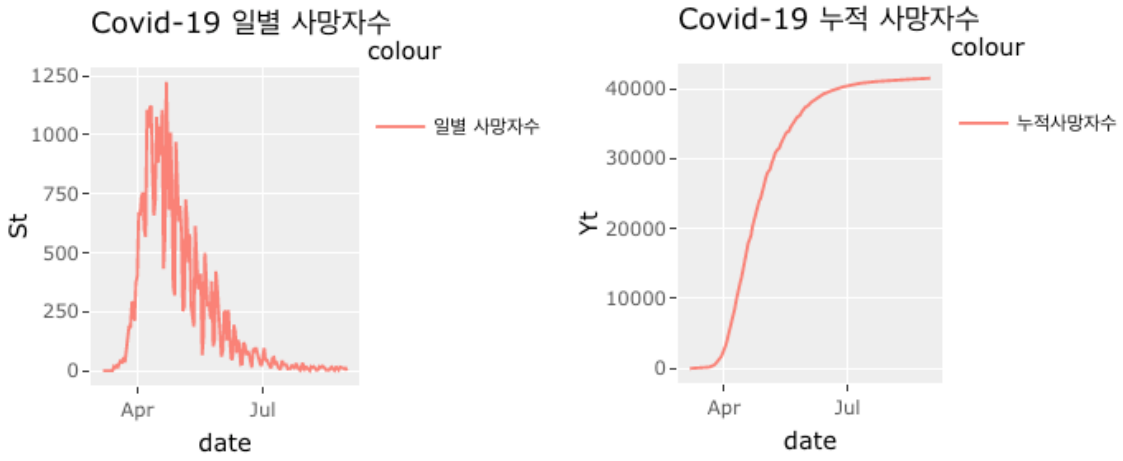


각 모형에 대한 QQ-plot 을 그려보았을 때, Gumbel 모형에서 가장 직선을 잘 따르는 것을 확인할 수 있다. Bass 모형도 직선과 가까운 형태를 보이지만, 1M-DRAM 자료가 censored 자료인 것을 고려했을 때, 상대오차 값이 양수가 나온 Gumbel 모형이 수요 예측에 더 적절할 것이다.

<Part B> 영국의 COVID-19 사망자 & HIV/AIDS 확산 예측

1. 영국 일별 Covid-19 사망자수 자료 (2020.3.7 - 2020.08.31)

a) 위 사이트에서 다운받은 엑셀자료를 이용하여 최초 사망 발생한 날부터 8 월 31 일까지 일별 신규 Covid-19 사망자수 S(t) 및 누적 사망자수 자료 Y(t)에 대한 시계열 도표를 그려보시오.



5 월의 사망자 수가 가장 컸고 그 이후로는 감소하는 경향을 보인다.

b) 최초 사망자가 발생한 날부터 초기 n일까지 일별 Covid-19 사망자수 자료 St ,t = 1, 2, 3, … ,n를 이용하여 장차 영국 내 총 감염자수(m)을 추정하려 한다.

 # N = 20, 30, 50 일 경우 세가지 모형을 이용한 해당 모수(p,q,m)들을 각각 추정하고 추정된 m 값과 최신 m 값(ex: 8 월 31 일까지 누적 사망자수)과 비교한 상대 오차 (100*(m^-m)/m)값을 구하시오.

N	Bass (m,p,q)	상대오차
20	2.778e+32, 3.78e+28, 3.78e+28	inf
30	1.239e+04, 3.539e-04, 0.2726	-70
50	2.724e+04, 2.371e-03, 0.1466	-34

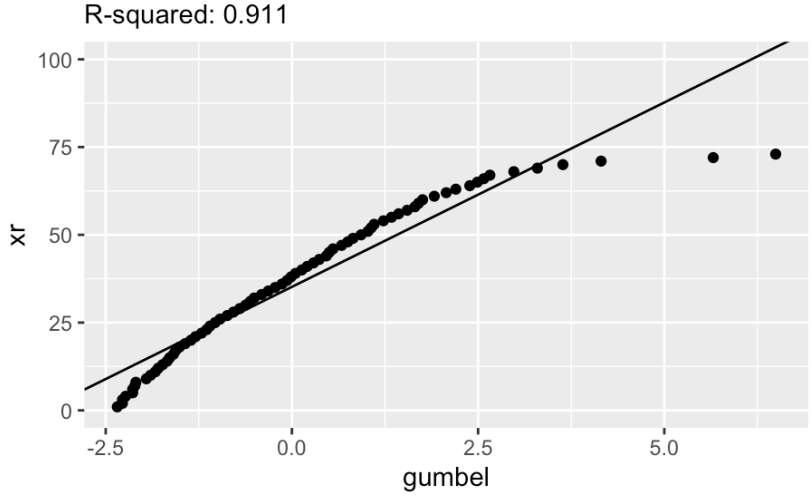
N	Logisitc(m,q) ; p=0	상대오차
20	-598.83, 0.204	-101.443
30	12065.71, 0.277	-70.925
50	26656.39, 0.158	-35.766

N	Gumbel(m,q) ; p=0	상대오차
20	4.0751, -0.0749	-99.99
30	135492.3, 0.0529	226.495
50	33872.14, 0.0776	-18.378

상대오차의 절대값이 가장 작은 Gumbel (N=50) Model 이 가장 좋은 모델이다.

c) 여러 추정 방법(ex: OLS, Q-Q plot 추정, 베이지안)을 이용한 m 추정값의 정확도를 비교해 보고 각 방법의 장단점을 기술하시오.

Gumbel (N=50) 모델의 Q-Q Plot 을 그려본 결과는 다음과 같다.



M 값이 너무 작아 gumbel : -log(-log(ur))로 하면, 114 개의 데이터가 누락된다는 단점이 있다.

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

N	Bass (m,p,q)	상대오차
20	12269.48, 0.005, 0.22	inf
30	19930.28, 0.005, 0.16	-65
50	28697.56, 0.0071, 0.092	-46

N	Logisitc (m,q) ; p=0	상대오차
20	10962.23, 0.257	-68.972
30	18815.09, 0.186	-46.745
50	27479.92, 0.118	-22.219

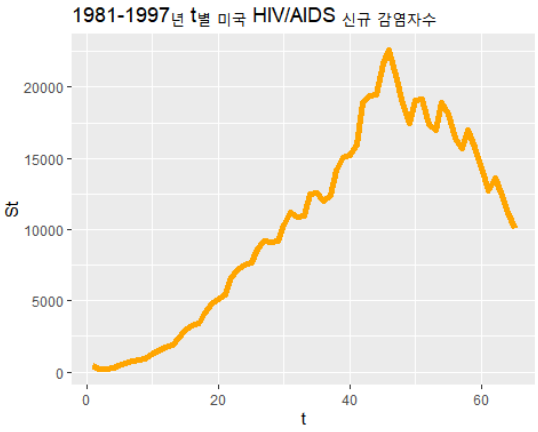
N	Gumbel (m,q) ; p=0	상대오차
20	23567.10, 0.086	-33.294
30	26598.35, 0.082	-24.714
50	30556.22, 0.069	-13.512

상대오차의 절대값이 가장 작은 Gumbel (N=50) Model 이 가장 좋은 모델이다.

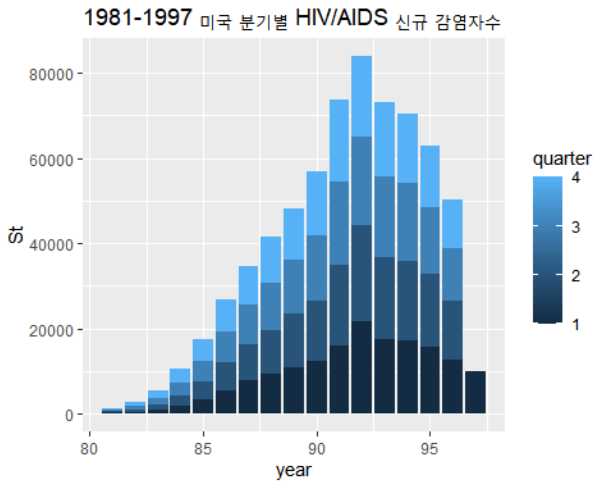
2. 미국 분기별 HIV/AIDS-감염자 자료 (1981-1997)

a) 분기별 HIV/AIDS 감염자자료(S(t))에 대한 시계열도표를 그려보시오.

S(t) : 신규 에이즈 감염자수



t 가 45-47 1992 년도 신규 에이즈 감염자수가 2 만명을 넘었고 1992 년 2 분기(t=46)에 최고점에 도달했다. 그 후부터는 감염자수가 현저하게 감소하는 추세임을 볼 수 있다.



분기별(1-4 분기) 신규 감염자수를 시계열도표로 나타냈다. 전반적으로 매년 총 신규 감염자수에 변화가 있어도 각 분기별 신규 감염자수 비율은 균일하다. 이를 통해 한 해 동안 각 분기마다 비슷한 비율로 감염자가 발생함을 알 수 있다. (분기에 관계없이 비슷하게 발생한다.)

b) 분기별 HIV/AIDS 감염자자료 S(t), t=1, 2, …, n 을 이용하여 장차 미국 내 총 감염자 수(m)을 추정하려 한다. n=20,40,50 일 경우 세 가지 모형의 해당 모수(p,q,m) 들을 각각 추정하고 추정값들을 최신 m 값(2012 누적 감염자수)과 비교한 상대오차를 구하고 그 의미를 설명하시오.

(실제 m 값 = 2012 년 미국 HIV/AIDS 누적 감염자수 = 1279443)

n	Bass	Logistic	Gumbel
	$(\hat{m}, \hat{p}, \hat{q})$	$(\hat{m}, \hat{q})$	$(\hat{m}, \hat{q})$
20	(116137, 0.00191, 0.214) 상대오차 = -90.923	(88572, 0.2456) 상대오차 = -93.077	(784235, 0.0503) 상대오차 = -38.705
40	(450452, 0.002, 0.12) 상대오차 = -64.793	(403336, 0.1366) 상대오차 = -68.476	(886444, 0.0461) 상대오차 = -30.716
65	(786740, 0.001525, 0.0965) 상대오차 = -38.509	(773275, 0.1033) 상대오차 = -39.562	(1000860, 0.0483) 상대오차 = -21.774

c) 최적모형 선택

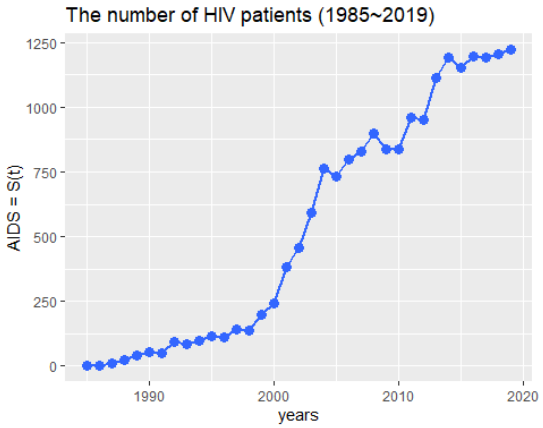
n이 20, 40, 65인 경우 모두 Bass와 Logistic모형에 비해 Gumbel 모형의 상대오차가 작다. 특히 n = 20일 때 Bass와 Logistic모형의 상대오차는 90이 넘지만 Gumbel 모형의 상대오차는 현저히 낮다. 즉, n이 작아도 Gumbel 모형은 낮은 상대오차를 보여준다.

n이 커질수록 모형의 정확도는 올라가며 n = 65일 때 세 모형의 추정한 m값을 비교해보면 Gumbel 모형의 추정값이 1000860으로 실제 m값인 1279443에 제일 근접한다.

따라서 미국 신규 AIDS/HIV 감염자수를 예측할 때 가장 적합한 모델은 Gumbel 모형이다.

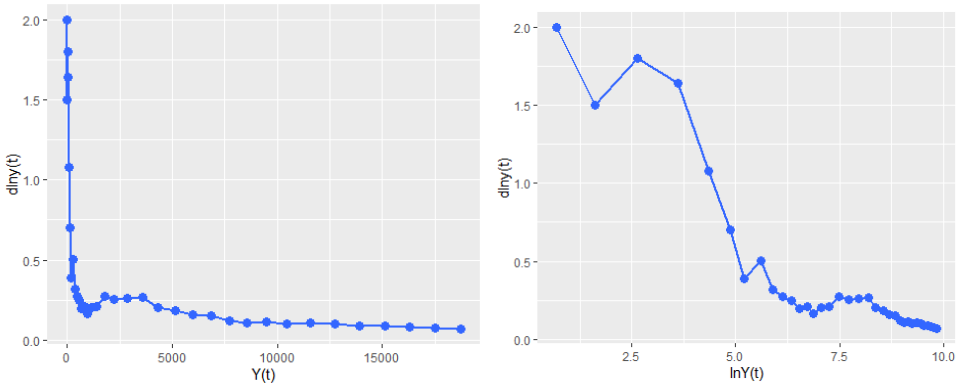
3. 한국 연도별 HIV/AIDS 감염현황자료 (1985-2019)

a) 한국 HIV 감염자자료 (1985-2019년, 명)를 이용하여 HIV 감염자수 $S(t)$, $t \leq 2019$의 시계열 도표를 그리시오.



국내 HIV 감염인의 숫자는 2000년까지는 완만하게 증가하다가 그 이후부터 가파르게 증가하는 추세를 보인다. 2015년부터는 증가추세가 약화된 것으로 보인다.

b) (dlnYt, Yt), (dlnYt, lnYt) 의 산점도를 각각 그리고, 각 경우 선형모형이 적절한지 검토하고 적절한 OLS 방법으로 (m,q) 를 추정하시오.



각 산점도에서 선형성을 찾기 힘들었기 때문에 dlny(t)와 y(t)를 직접적으로 이용하는 식이 아닌, s(t)와 y(t)를 이용하는 식을 사용해서 회귀 직선을 추정하였다.

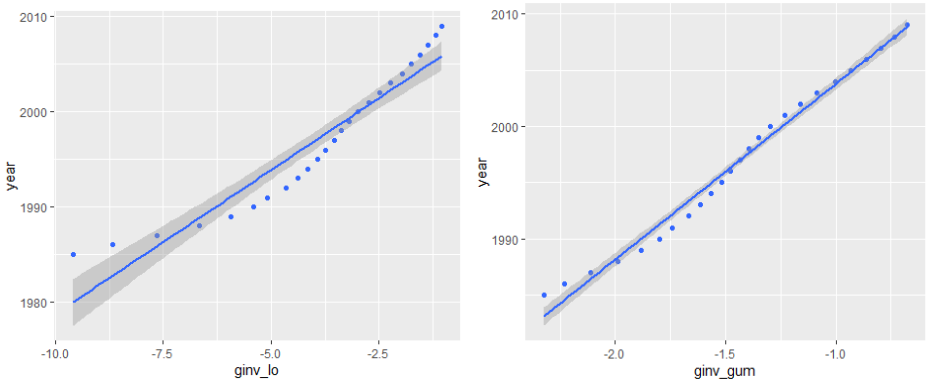
	Regression	Adj R-squared	(q, m)
Logistic	S(t) = 0.165y(t-1) - 5.6e-6 $y(t-1)^2$	0.9855	(0.165, 29400)
Gumbel	S(t) = 0.66y(t-1) - 0.06y(t) lny(t-1)	0.9928	(0.06, 54871)

Adjusted R-squared 를 비교한 결과 Gumbel 회귀모델의 값이 더 높다는 것을 알 수 있었다. 즉, 국내 HIV 감염인 현황을 예측하는 데에 더 적절한 모델은 Gumbel 모델이다.

c) 위에서 추정한 m을 기반으로 Logistic 및 Gumbel Q-Q plot 그리기 & (μ, σ) 추정

(단, 자료는 1995-2009년)

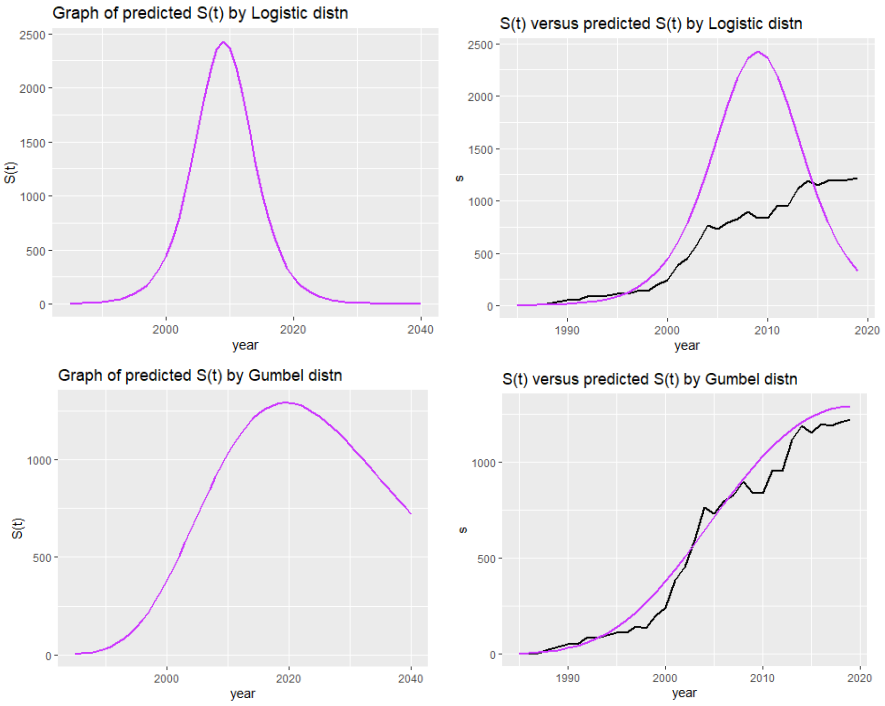
 # Logistic Distribution **# Gumbel Distribution**



	Logistic	Gumbel
Adj R-squared	0.9076	0.986
μ	2009.02	2019.434
σ	3.026	15.636

Regression 때와 마찬가지로, Q-Q Plot의 Adjusted R-squared를 비교했을 때 Gumbel분포의 값이 더 높다. 즉, Gumbel 분포가 1985년부터 2009년까지의 데이터를 더 잘 설명한다고 할 수 있다.

d) 추정된 모수값 (m, μ, σ) 을 이용하여 S(t)의 예측값을 추정하고 (1985~2040년) 추정값과 실제 S(t)값 겹쳐 그린 후 예측값의 의미를 설명하시오.

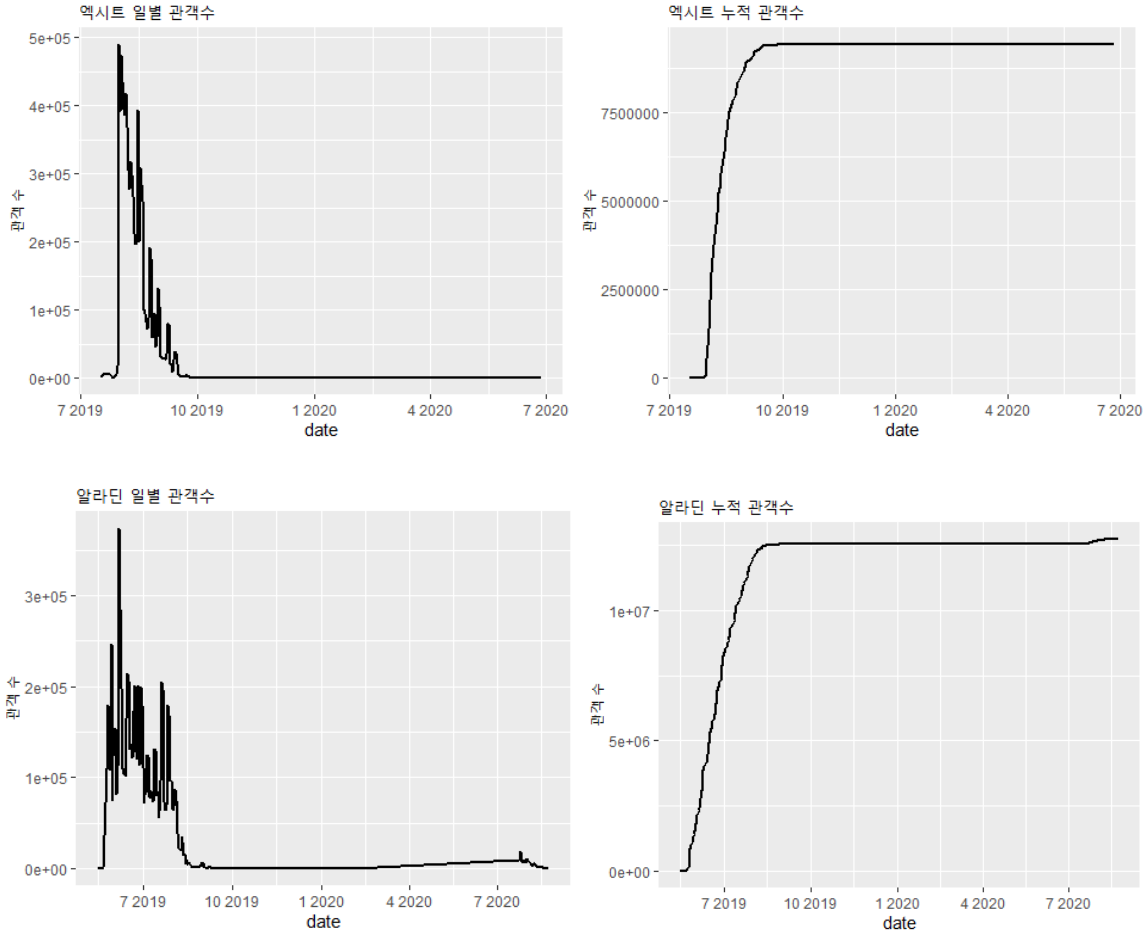


Logistic 분포로 구한 예측값은 약 2010년에 최댓값을 기록하고 점점 감소함을 볼 수 있다. 이를 실제 관측값과 비교해보면, logistic 분포는 S(t) 값을 예측하기에 적절하지 않음을 알 수 있다.

그러나 Gumbel 분포로 구한 예측값은 약 2020년 최댓값을 기록하고 그 이후 점점 감소하는 경향을 보임을 알 수 있다. 이를 실제 관측값과 비교해보면 꽤 잘 맞다는 것을 볼 수 있고 결론적으로 Gumbel 분포가 적절한 분포임을 확인할 수 있다.

<Part C> 영화 흥행 예측

1) 일별 관객수S(t) 및 누적 관객수Y(t) 시계열 도표를 그리시오.



엑시트는 약 2019년 9월을 기점으로 증가하지 않았고, 알라딘은 약 2019년 8월을 기점으로 증가하지 않았다.

2) 아래 4가지 확산 모형과 개봉 후 1주, 2주 및 4주 간 흥행 자료를 이용하여 총 관객수(m)을 추정한 후 이를 실제 총 관객수 m값과 비교한 상대오차 값을 구하여 최적 예측 모형을 찾아보시오.

엑시트

총 관객수 추정 (실제 m : 9426161)				
	bass	logisic	gumbel	Exponent
Week1	15303	15071	15264	21640
Week2	2355667	2297368	2648178	-275878
Week4	6794898	6640789	7073915	-5423895

상대오차				
	bass	logisic	gumbel	Exponent
Week1	-99.84	-99.84	-99.84	-99.77
Week2	-75.01	-75.63	-71.91	-102.93
Week4	-27.91	-29.55	-24.95	-157.54

알라딘

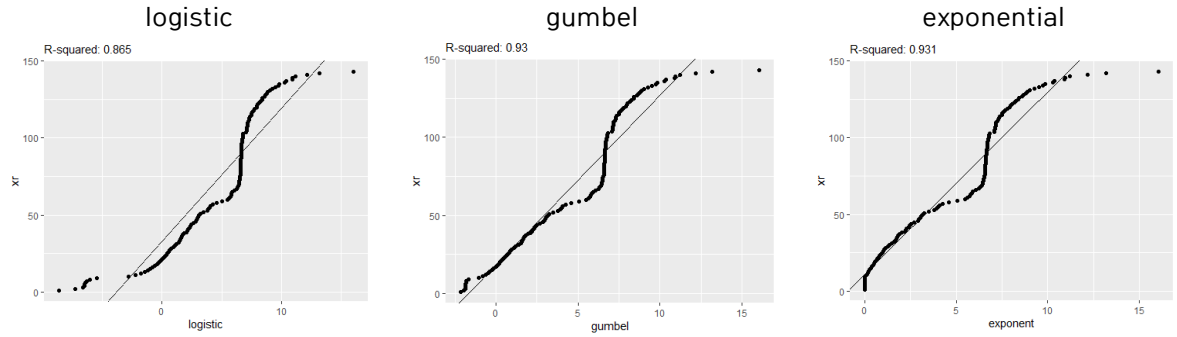
총 관객수 추정 (실제 m : 12723775)				
	bass	logisic	gumbel	Exponent
Week1	372106	341970	562974	-20401
Week2	1528417	1487495	1640016	-675373
Week4	8088579	5704853	7697641	-1775874

상대오차				
	bass	logisic	gumbel	Exponent
Week1	-97.08	-97.31	-95.58	-100.16
Week2	-87.99	-88.31	-87.11	-105.31
Week4	-36.43	-55.16	-39.50	-113.96

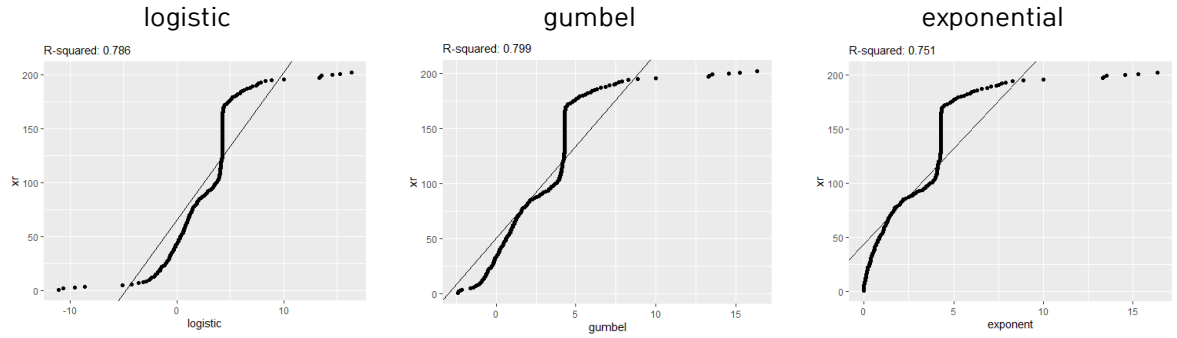
엑시트는 자료의 개수가 가장 많을 때에는 gumbel모델의 상대오차가 제일 작았고, 알라딘은 bass 모델의 상대오차가 제일 작았다. Exponential 모델은 모든 데이터에서 다른 분포들에 비해 상당히 큰 상대오차를 갖게 되므로 적절하지 않은 분포라고 판단할 수 있다.

3) 실제 총 관객수 m을 이용하여 해당 모형의 Q-Q plot을 그려보고 해당 모형이 적절한지 검토.

엑시트



알라딘



Q-Q plot을 그려본 결과 엑시트는 gumbel과 exponential 모형이, 알라딘의 경우 gumbel 모형이 가장 적절한 것을 알 수 있었다.

< Part D > : R shiny Link : http://127.0.0.1:4609/