베이지안 종합시험

- 뎅기열 확산 사례수 데이터를 이용한 분위 회귀 모형 및 예측-

학번	학과	이름
182STG12	통계학과	오혜윤

목차

분석 목적 및 방법론

*뎅기열 확산 예측

뎅기열은 세계의 열대 및 아열대 지역에서 발생하는 모기 매개 질병이다. 경증의 경우, 증상은 독감과 비슷하다. 발열, 발진, 근육 및 관절통뿐만 아니라 심한 경우에는 뎅기열로 심한 출혈, 저혈압, 사망까지 유발할 수 있다. 또한, 뎅기열은 모기에 의해 운반되기 때문에, 뎅기열의 전달은 온도와 강수량과 같은 기후 변수와 관련이 있다. 뎅기열과 기후와의 관계는 설명하기 복잡하나 과학자들은 전 세계적으로 기후 변화가 공중 보건에 중요한 영향을 미치는 변화를 일으킬 것이라고 주장한다.

최근 뎅기열이 전세계적으로 퍼지고 있다. 동남아시아 및 태평양 섬에 가장 널리 퍼졌다. 요즘 라틴 아메리카에서는 연간 약 5 억 건의 사례가 발생하고 있다고 한다.

Driven Data 에서 제공한 뎅기열 확산건수 데이터를 활용하여 온도, 강수량, 초목 등의 환경 변수에 따라 매주 뎅기열 사례 수를 예측할 수 있다. 이러한 기후와 뎅기열의 관계에 대한 이해는 전세계의 생명을 위협하는 전염병에 대처할 수 있는 방안을 마련할 것이다.

* 베이지안 예측, 분위회귀모형 및 변수 선택

베이지안 방법론을 통하여 뎅기열 확산건수를 예측하고, 뎅기열 확산에 영향을 미치는 변수들에 대해 알아본다. 뎅기열 확산건수는 다양하게 분포되어 있으나, 대부분의 확산건수가 앞쪽에 몰려있는 skewed 형태를 띄기 때문에 이 때 분위 회귀 모형을 도입하여 확산건수의 분위 별 영향 변수를 각각 따로 알아보도록 한다. 또한, 이렇게 영향을 미치는 변수들을 알아본 뒤, 변수 선택을 통해 변수들의 사후 확률을 알아보고 적합한 예측 모델의 구조를 알아본다.

평균 회귀 모형에서는 y의 조건부 평균에 초점을 맞추어, $\mathrm{Y} = x_i{}' \beta + \epsilon_i,\ \epsilon_i \sim iid\ \ \mathrm{N}\ (0,\sigma_2)$ 식을 기반으로 예측을 전개한다. 하지만, 이러한 조건은 만족되지 않을 때가 많으며, 에러가 증가한다면, 데이터를 정확하게 설명하기 어려워진다. 따라서, $x = (x_1, ..., x_k)$ 일 때, $\mathrm{P}\ (\mathrm{Y} \leq \mathrm{Q}_{\ p}(\mathrm{Y}\ |x\ )) = p\ (0 < p < 1.)$ 를 만족시키는 $\mathrm{Q}_p\ (\mathrm{Y}\ |x\ )$ y의 p분위수로 하여 $\mathrm{Q}_{\ p}(\mathrm{Y}\ |x\ ) = x'\beta_p$ 을 도출할 수 있다.

β_p 을 는 체크 손실함수를 최소화하는 추정치이며, 베이지안 추론을 위한 우도함수를 도출하기 위하여 ϵ_i에 대하여 ALD 모형을 도입한다.

이후, 깁스변수선택(GVS) 기법을 사용하여 뎅기열 확산 데이터에서 반응변수인 뎅기열 확산건수에 유의한 영향을 미치는 설명변수들을 찾아보도록 한다. 또한, 선택된 변수들의 계수 예측값을 이용하여 예측 모델을 정의하고 정확도를 계산해본다.

Part 1 Data 전처리 과정

1. Data 설명

전체 데이터는 위치, 날짜, 기후 및 뎅기열 확산건수 데이터로 이루어져 있다. 총 1456개의 관측치를 포함한다.

1) train 데이터는 24개의 변수를 포함한다. y는 total_cases (뎅기열 확산건수)이며, 기타 변수들은 온도, 습도, 강수량 등으로 이루어져 있다.
2) 강수량 및 온도 데이터는 각각 관측소가 달라, 중복되어 있는 경우가 많다. 상관관계를 고려하여 전처리 과정에서 제거하도록 한다.
3) 결측치가 상당히 많이 존재하여, 결측치를 대체하는 여러가지 방법을 도입하여 대체하기로 한다.

* train 데이터 예시*

변수명	상세 설명	변수명	상세 설명
Total_cases	뎅기열 확산건수	reanalysis_max_air_temp_k	최대 공기온도(NCEP)
city	도시명	reanalysis_min_air_temp_k	최소 공기 온도(NCEP)
year	년도	reanalysis_precip_amt_kg_per_m2	총 강수량(NCEP)
weekofyear	주	reanalysis_relative_humidity_percent	평균 상대 습도(NCEP)
week_start_date	날짜	reanalysis_sat_precip_amt_mm	총 강수량(NCEP)
ndvi_ne	위치 좌표(북동쪽)	reanalysis_specific_humidity_g_per_kg	평균 비 습도(NCEP)
ndvi_nw	위치 좌표(북서쪽)	reanalysis_tdtr_k	주간 온도 범위 (NCEP)
ndvi_se	위치 좌표(남동쪽)	station_avg_temp_c	평균 온도(GHCN)
ndvi_sw	위치 좌표(남서쪽)	station_diur_temp_rng_c	주간 온도 범위(GHCN)
precipitation_amt_mm	총 강수량(PERSIANN)	station_max_temp_c	최대 온도(GHCN)
reanalysis_air_temp_k	평균 대기 온도 (NCEP)	station_min_temp_c	최소 온도(GHCN)
reanalysis_avg_temp_k	평균 기온 (NCEP)	station_precip_mm	총 강수량(GHCN)
reanalysis_dew_point_temp_k	평균 이슬점 온도 (NCEP)		

2. 전처리 과정 (EDA)

먼저, 분석을 행하기에 앞서 데이터 정제 및 결측치 처리, 이상치 처리를 진행한다. 이후, plot을 그려보고 변수 변환을 시도한다.

① 이상점 및 결측치 처리

다음 그림은 각 변수 별 결측치를 퍼센트로 나타낸 것이다. 변수 별로 상당히 많은 결측치가 존재하는 것을 볼 수 있다.

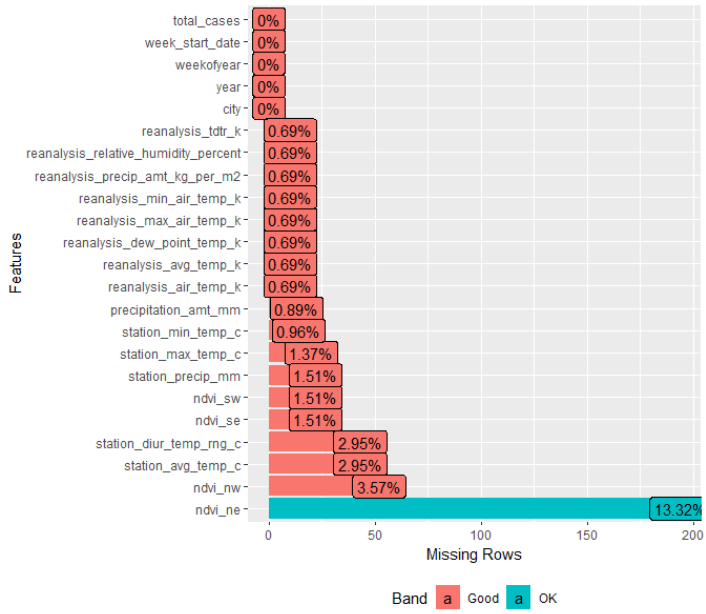


그림 1: 변수 별 결측치 그래프

* 이상점 (Outlier) 및 결측치 (Missing Value) 처리

- 관측치의 10개 이상의 설명변수 값이 결측일 경우 해당 관측치 제외 (10개 자료점 제외) - 결측치를 평균으로 대체

- 상세 처리 과정

Step 1 : dataset에 대한 이상점들을 1차 제거
(관측치들을 모든 변수를 고려한 k개의 이웃으로 나누어 관측치별 outlier score을 계산 후 score 기준으로 상위 5개의 값을 제외함)

Step 2 : 이상점들을 제거한 dataset을 통하여 평균으로 결측치를 추정

Step 3 : Step2에서 분석에 사용할 dataset에 대하여 이상점들을 2차 제거 (2개의 값 제거)

Step 4 : 최종적으로 1439개 관측치가 있는 dataset을 분석에 사용

- 결측치 대체 방법.

Step 2에서 결측치 대체하기 위해 K-NN clustering, Mean, Median, RPART package 총 3가지 방법을 사용하였다.

그 중 평균으로 대체하는 것이 MSE 값이 가장 작았기 때문에 결측치 처리방법으로 선택하였다.

precipitation_amt_mm	**K-NN**	**Mean**	**Median**	**RPART**
MSE	7.8521	7.5199	1862.8818	30.7543

② Correlation

관측소가 다른 각각 precipitation_amt_mm(총 강수량)와 reanalysis_sat_precip_amt_mm(총 강수량)은 상관계수가 1이며, reanalysis_dew_point_temp_k(평균 이슬점 온도)와 reanalysis_specific_humidity_g_per_kg(평균 상대 습도)의 상관계수가 0.997이므로 하나의 변수만 취하기로 한다.

또한, reanalysis_air_temp_k(평균 대기 온도)와 reanalysis_avg_temp_k(평균 기온)도 상관계수가 0.9016로 상당히 높기 때문에 두 변수 중 하나만 취하기로 한다. precipitation_amt_mm(총 강수량)과 reanalysis_specific_humidity_g_per_kg(평균 상대 습도), reanalysis_avg_temp_k(평균 기온)으로 진행한다.

③ 변수 변환

변수 별 frequency density plot을 그려, 변환을 할 것인지 결정한다. skewed되었을 경우, log 변환을 고려한다.
station_precip_mm, station_diur_temp_rng_c, reanalysis_tdtr_k, precipitation_amt_mm, reanalysis_precip_amt_kg_per_m2, total_cases 총 6가지 변수에 대해 log 변환을 하였다.
그리고 year와 city 변수에 대해서는 factor화시켰다.

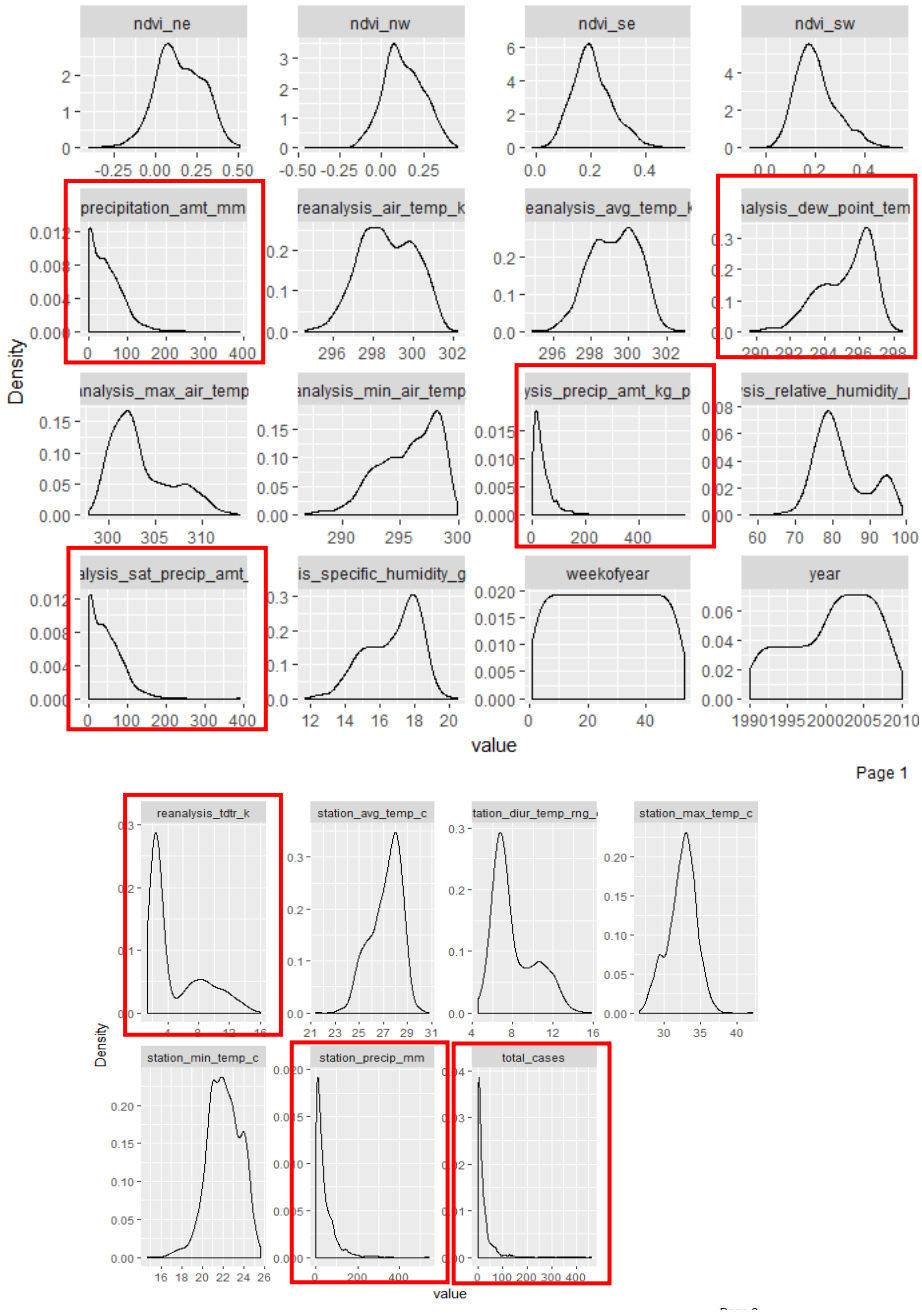


그림 2 : 변수 별 density plot

Part 2 모델링 및 결과

1. EDA

*Marginal Plot

① City 별 total_cases (확산건수) 그래프

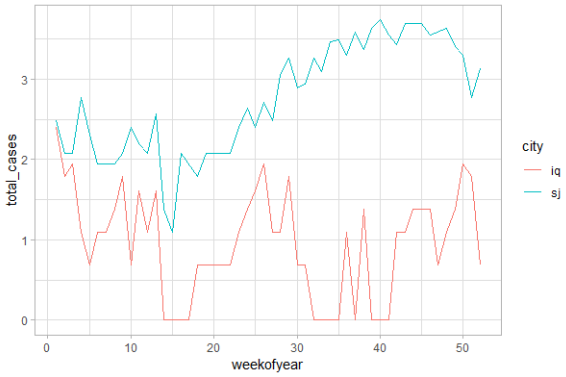


그림 3 : 2003 년

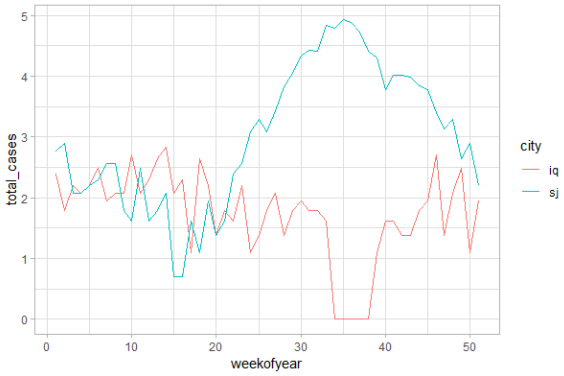


그림 4 : 2005 년

도시 sj는 1년 중 3분기, 4분기에 갑자기 뎅기열 확산건수가 증가한다. 대부분의 년도에서 위와 같은 그래프를 보이는 것을 확인하였다. 하지만, 도시 iq는 sj보다 훨씬 낮은 확산건수를 보이며, 분기 별 뚜렷한 추세가 보이지 않는다.

이 그래프를 통하여, 년도 별 추세를 알아볼 수 있으며, 년도 별 추세가 있다는 것은 기후와 밀접한 관련이 있다는 것을 의미하기도 한다.

② 평균 습도 vs 확산건수 그래프와 ② 평균 기온 vs 확산건수 그래프와 (도시 sj 한정)

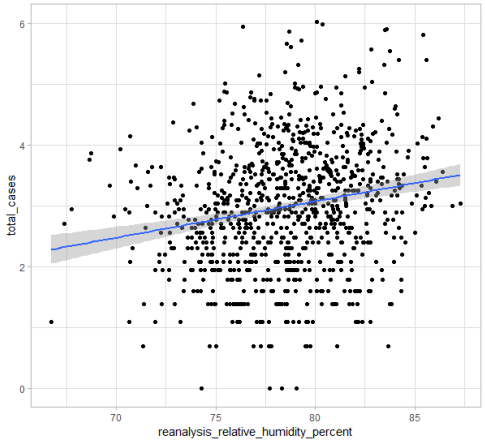


그림 5 : 도시 sj 의 평균 습도 vs 확산건수

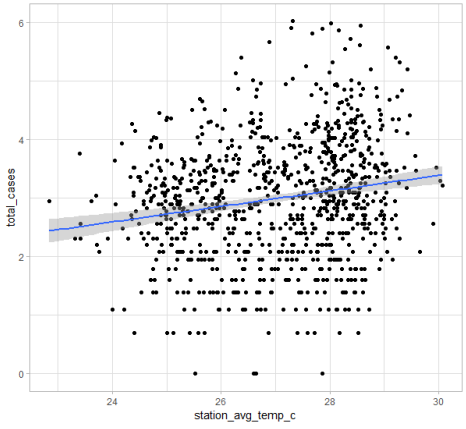


그림 6: 도시 sj 의 평균 기온 vs 확산건수

그래프를 그려본 결과, 각각 습도와 기온이 확산건수에 대해 약간의 linear한 관계를 보인다.

2. 모델링

Training set과 Test set을 8:2로 랜덤 추출하여 진행한다. 분위회귀모형을 통해, 영향력있는 변수들을 살펴보고, 마지막은 여러가지 모델링의 mse를 비교하여 예측한다.

R-square: 58%

Coefficients:	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	160.110120	36.043338	4.442	9.79e-06	***
citysj	-0.395673	0.316559	-1.250	0.211589	
year1991	1.158444	0.201721	5.743	1.20e-08	***
year1992	1.310357	0.204715	6.401	2.27e-10	***
year1993	0.601389	0.198228	3.034	0.002470	**
year1994	1.732993	0.210020	8.252	4.35e-16	***
year1995	0.292780	0.200141	1.463	0.143784	
year1996	0.153558	0.202711	0.758	0.448896	
year1997	0.443153	0.202963	2.183	0.029212	*
year1998	1.680837	0.210679	7.978	3.64e-15	***
year1999	0.693332	0.212892	3.257	0.001161	**
year2000	-0.997775	0.193710	-5.151	3.06e-07	***
year2001	-0.397521	0.188992	-2.103	0.035655	*
year2002	0.087463	0.192100	0.455	0.648980	
year2003	-0.163219	0.191781	-0.851	0.394913	
year2004	0.230994	0.196923	1.173	0.241037	
year2005	0.108079	0.196614	0.550	0.582634	
year2006	-0.182284	0.196778	-0.926	0.354467	
year2007	0.416432	0.193720	2.150	0.031796	*
year2008	0.258709	0.210704	1.228	0.219769	
year2009	-0.048516	0.221165	-0.219	0.826405	
year2010	0.309367	0.255545	1.211	0.226297	
weekofyear	0.007463	0.001940	3.846	0.000127	***
ndvi_ne	0.033951	0.333400	0.102	0.918907	
ndvi_nw	1.118168	0.399312	2.800	0.005194	**
ndvi_se	0.421106	0.594085	0.709	0.478576	
ndvi_sw	-0.596626	0.576204	-1.035	0.300686	
precipitation_amt_mm	0.001246	0.023382	0.053	0.957507	
reanalysis_avg_temp_k	-0.421295	0.136495	-3.087	0.002075	**
reanalysis_max_air_temp_k	-0.084424	0.036020	-2.344	0.019261	*
reanalysis_min_air_temp_k	-0.008051	0.042800	-0.188	0.850835	
reanalysis_precip_amt_kg_per_m2	-0.024833	0.046009	-0.540	0.589485	
reanalysis_relative_humidity_percent	-0.164485	0.028196	-5.834	7.09e-09	***
reanalysis_specific_humidity_g_per_kg	0.899500	0.134981	6.664	4.17e-11	***
reanalysis_tdtr_k	0.034119	0.296632	0.115	0.908450	
station_avg_temp_c	-0.197069	0.053904	-3.656	0.000268	***
station_diur_temp_rng_c	0.198309	0.312470	0.635	0.525786	
station_max_temp_c	-0.010503	0.032574	-0.322	0.747176	
station_min_temp_c	-0.007567	0.036712	-0.206	0.836736	

```
station_precip_mm                    0.013272   0.024538   0.541 0.588716
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 0.8068 on 1122 degrees of freedom
Multiple R-squared:  0.5782, Adjusted R-squared:  0.5635
F-statistic: 39.43 on 39 and 1122 DF,  p-value: < 2.2e-16
```

유의한 변수로 year(년도), weekofyear(주), ndvi_nw(좌표 북서), reanalysis_avg_temp_k(평균 온도), reanalysis_max_air_temp_k(최대 온도), precipitation_amt_mm(평균 강수량), reanalysis_min_air_temp_k(최소 공기 온도) ,reanalysis_relative_humidity_percent(평균 상대 습도), reanalysis_specific_humidity_g_per_kg(평균 비 습도), station_avg_temp_c(평균 온도)가 선택되었다.

또한, R-square 58%로 다소 높은 설명력을 보였다.

나중에 있을 모델링 비교를 위하여, Training set과 Test set으로 나누어 예측을 시행한 결과,MSE = 0.7606로 측정되었다.

① 평균 회귀 분석-lasso

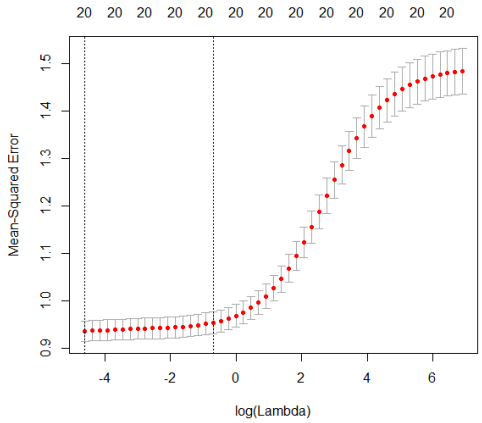


그림 7: Lasso Lambda

최적의 lambda = 0.0045이고, Training set과 Test set으로 나누어 예측을 시행한 결과, MSE = 0.9504로 측정되었다.

② 평균 회귀 분석-ridge

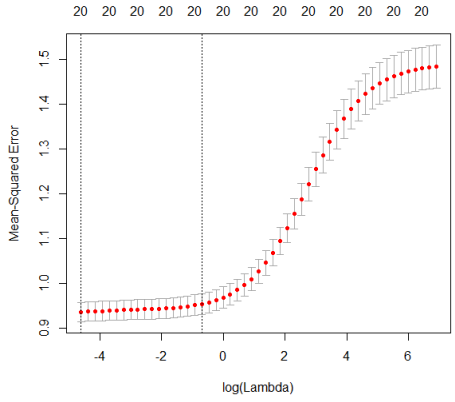


그림 8: Ridge Lambda

최적의 lambda = 0.1이고, Training set과 Test set으로 나누어 예측을 시행한 결과, MSE = 0.9497로 측정되었다.

다음은 Linear Regression에 대한 MSE 비교 표이다. OLS 방법이 0.7606로 MSE가 가장 작은 것으로 나타났다.

	OLS	LASSO	RIDGE
MSE	0.7606	0.9504	0.9497

③ 분위 회귀 분석

앞서 최종 데이터셋을 Linear Regression에 적합시킨 결과, Linear Regression이 적절하다는 사실을 발견하였다. 따라서, 분위 회귀 모형을 도입하여 분위 별 영향력을 보기로 하였다. Total_cases (확산건수)의 경우 다소 0 근방에 모두 모여 있어 분위 별로 자세하게 분석하기 위해 분위회귀모형을 도입하였다.

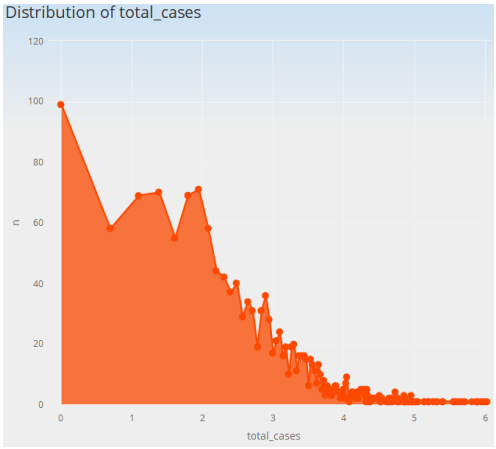


그림 9: Total_cases(확산건수) Density plot

선형 회귀모형을 적합시켜 얻은 계수 추정치를 정규 사전분포의 평균으로 사용하고 추정치의 분산에 100을 곱하여 사전분산으로 사용하였다. 그리고 0-1 기법을 사용하여 우도함수를 직접 지정하기 위하여 베르누이 분포를 따르는 가상 자료 Ones를 DataList 에서 제공하였다. 끝으로, 분위수에 대하여 분위회귀모형 회귀계수 추정치를 구한 후 (95% 사후구간 포함) 각 설명변수에 대하여 분위수에 따른 추정치의 변화를 그림으로 그려보았다. 각 과정은 iteration 30000번, Chain 수 3, nAdapt=1000; nUpdate=10000 으로 설정하였다.

예시로 beta 21 추정치의 경로 그림 결과는 다음과 같다.

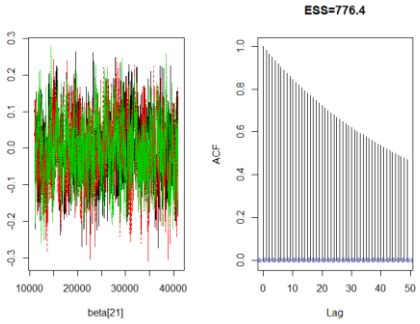


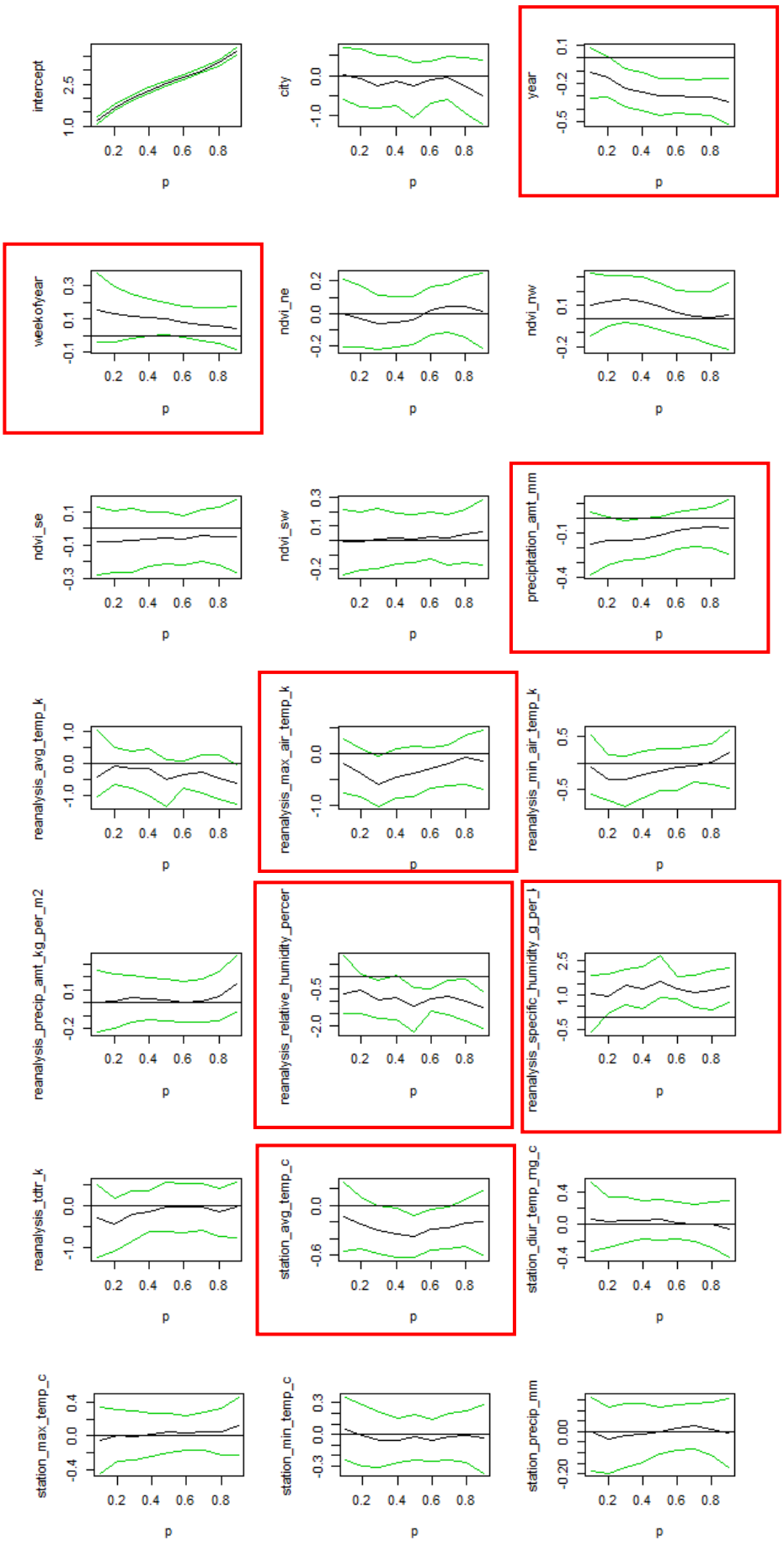
그림 10: beta 21 경로그림

scaling을 하고 진행하였지만, 다소 자기상관성을 보이는 것으로 판단된다. 경로그림은 잘 수렴하는 것으로 판단되었다.

beta21 뿐만 아니라 다른 추정된 beta들이 모두 ACF 그래프가 천천히 감소하는 것을 확인하였다. 따라서,

자기상관성이 제거되지 않았음을 확인할 수 있었다. 특히, beta 10, beta11에 대해서는 경로그림 또한 잘 수렴하지 않는 형태를 보였다. beta10, beta11은 모두 온도에 관한 변수로 모든 변수에 대해 상관관계가 다소 크기 때문이라고 생각하였다. 기후 데이터 간의 상관관계가 워낙 어쩔 수 없이 크기 때문에 표준화를 시켰는데도 불구하고 자기상관성이 크게 보임을 알 수 있었다.

분위수에 따라 추정된 분위회귀계수와 95% 신뢰구간을 그린 결과는 다음과 같다.



회귀계수 추정치가 분위수에 따라 달라지는 것을 확인할 수 있었다. 유의한 변수로 year(년도), weekofyear(주), precipitation_amt_mm(평균 강수량), reanalysis_max_air_temp_k(최대 온도), reanalysis_relative_humidity_percent(평균 상대 습도), reanalysis_specific_humidity_g_per_kg(평균 비 습도), station_avg_temp_c(평균 온도)가 선택되었다.

선택된 변수들은 앞서 ols로 선택된 유의한 변수들과 두 가지 변수를 제외하곤 모두 똑같았다.

먼저 year 변수는 모든 분위에서 음의 영향력을 미쳤고, 이것은 년도가 증가함에 따라, 뎅기열 확산건수가 감소한다는 것을 의미한다. 이러한 결과가 도출된 것은 현재 예전에 비하여 질병 관련 대처가 철저하게 이루어지기 때문이라고 생각하였다. 또한, 영향력의 크기는 분위수마다 달랐는데, 중하위분위에 비해 상위 분위에서 음의 영향력이 더 크게 나타났다. 년도가 증가하면 뎅기열 확산건수가 많은 곳이 적은 곳에 비해 뎅기열 확산건수 하락 폭이 더 커짐을 의미한다.

weekofyear 변수는 모든 분위에서 양의 영향력을 미쳤고, 이것은 주가 증가함에 따라, 뎅기열 확산건수가 증가한다는 것을 의미한다. 즉, 3~4분기에 뎅기열 확산건수가 더 많을 것을 앞서 확인했던 결과와 비슷하게 도출해낼 수 있다. 또한, weekofyear 변수는 중하위 분위수에서 그 영향력이 더 큰 것으로 나타났다. 따라서 주가 증가하면 뎅기열 확산건수가 적은 곳이 많은 곳에 비해 뎅기열 확산건수 증가 폭이 더 커짐을 의미한다.

precipitation_amt_mm(평균 강수량) 변수는 음의 영향력이 분위수가 증가함에 따라 계속해서 감소하였고, 강수량이 증가하면, 뎅기열 확산건수가 감소하는 것을 의미한다.

reanalysis_max_air_temp_k 변수는 중간 분위수에서 서서히 음의 영향력이 감소하였고, 양 끝 분위수에서는 음의 영향력이 증가하는 형태를 보인다. 이것은 먼저 최대온도가 증가할수록 뎅기열 확산 건수는 감소하나, 확산건수가 극도로 크거나 작은 그룹은 중간 분위수 그룹에 비하여 뎅기열 확산건수 감소 폭이 증가함을 의미한다.

reanalysis_relative_humidity_percent 변수는 음의 영향력이 분위수가 증가함에 따라 계속해서 증가하였고, 상대습도가 증가할수록 뎅기열의 확산건수가 감소함을 의미한다. 또한, 상대 습도가 증가하면 뎅기열 확산건수가 많은 곳이 적은 곳에 비해 뎅기열 확산건수 하락 폭이 더 커짐을 의미한다.

마찬가지로, reanalysis_specific_humidity_g_per_kg 변수는 양의 영향력이 분위수가 증가함에 따라 다소 증가하는 형태를 보이고, station_avg_temp_c 변수는 음의 영향력이 증가하다 감소하는 형태를 보인다.

이러한 결과를 통하여, 년도, 상대습도, 강수량, 온도가 증가함에 따라 뎅기열 확산 건수가 감소함을 알 수 있었다. 무더운 온도에서 확산이 잘된다는 통상적인 결과와 사뭇 달랐다. 이는 전의 경로그림에서 볼 수 있었다시피, 온도를 나타내는 beta10, beta11에 대한 수렴이 제대로 이루어지지 않았기 때문에, 분위 회귀 모형에서도 제대로 결과가 도출되지 않은 것으로 생각하였다.

④ 변수 선택

가장 좋은 모델은 year(년도), reanalysis_avg_temp_k(평균 온도), reanalysis_tdtr_k(주간 온도 범위)만을 포함한 모델로, 사후확률은 0.3905로 측정되었다. 두번째로 높은 모델은 year(년도), reanalysis_min_air_temp_k(최소 온도)만을 포함한 모델로, 사후확률은 0.2177로 측정되었다. 모두 온도가 유의한 변수로 선택됨을 알 수 있었다. 그 외의 모델들은 사후확률이 낮아 의미가 없다고 판단하였다.

	intercept	city	year	weekofyear	ndvi_ne	ndvi_nw	ndvi_se	ndvi_sw	precipitation_amt_mm	reanalysis_avg_temp_k	reanalysis_max_air_temp_k	reanalysis_min_air_temp_k	reanalysis_precip_amt_kg_per_m2
162	1	0	1	0	0	0	0	0	0	1	0	0	0
117	1	0	1	0	0	0	0	0	0	0	0	1	0
223	1	0	1	0	0	0	0	0	0	1	1	0	0
11	1	0	0	0	0	0	0	0	0	1	0	0	0
128	1	0	1	0	0	0	0	0	0	0	0	1	0

reanalysis_relative_humidity_percent	reanalysis_specific_humidity_g_per_kg	reanalysis_tdtr_k	station_avg_temp_c	station_diur_temp_rng_c	station_max_temp_c	station_min_temp_c	station_precip_mm	prob
0	0	1	0	0	0	0	0	0.39048889
0	0	0	0	0	0	0	0	0.21768889
0	0	0	0	0	0	0	0	0.02573333
0	0	1	0	0	0	0	0	0.02351111
0	0	0	1	0	0	0	0	0.02182222

*참고

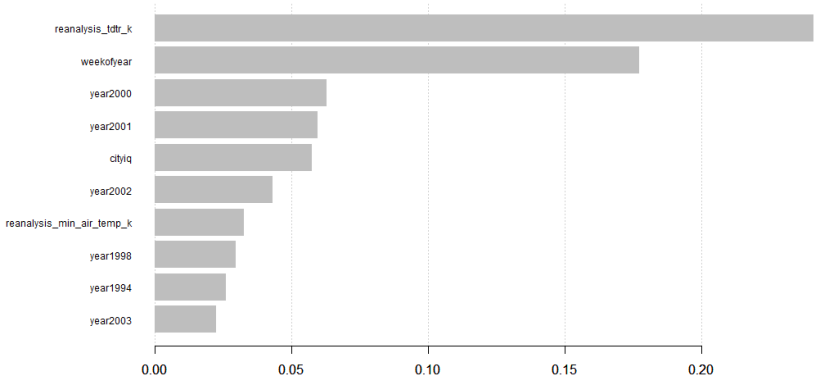
	포함 변수
stepAIC	city + year + weekofyear + ndvi_nw + reanalysis_avg_temp_k + reanalysis_max_air_temp_k + reanalysis_relative_humidity_percent + reanalysis_specific_humidity_g_per_kg + station_avg_temp_c
GVS	year+ reanalysis_avg_temp_k+ reanalysis_tdtr_k

⑤ Machine Learning Model

A. SVM

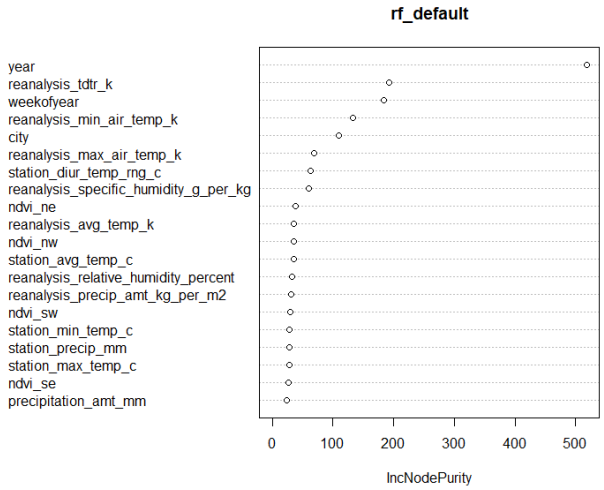
SVM 튜닝 결과, 최적의 gamma=0.01, cost = 10가 나왔다. 예측 MSE = 0.9374로 측정되었다.

B. XGboost



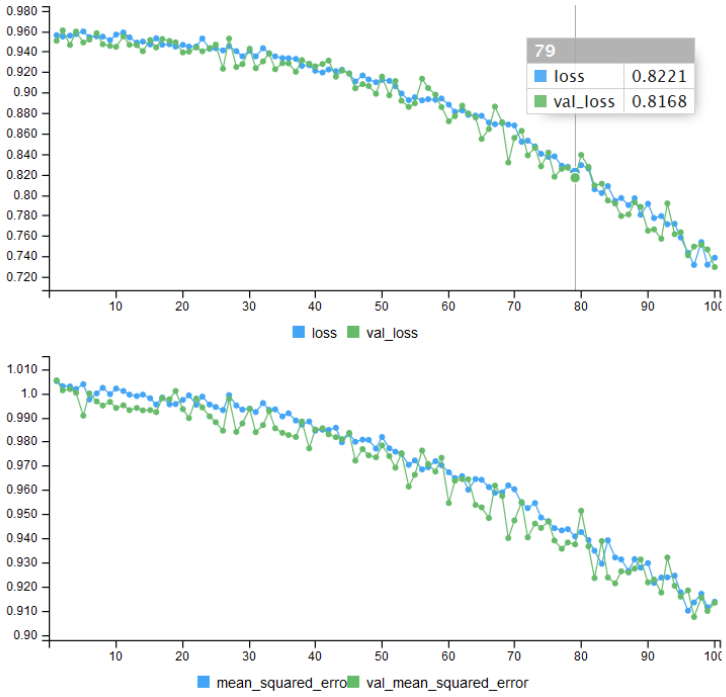
XGboost 결과, MSE= 0.2835로 측정되었다.
또한, 변수 중요도 플랏을 그려본 결과, reanalysis_tdtr_k, weekofyear, year 순으로 중요하다는 결과가 도출되었다.

C. RandomForest



랜덤포레스트 결과, MSE = 0.3707271로 측정되었다.
또한, 변수 중요도 플랏을 그려본 결과, year , reanalysis_tdtr_k, weekofyear 순으로 중요하다는 결과가 도출되었다.

D. Bayesian Machine Learning with keras



Bayesian ML 결과, MSE = 0.8945로 측정되었다.

*모델링 별 MSE 비교

	OLS	SVM	XGboost	RandomForest	Bayesian ML
MSE	0.7606	0.9374	0.2835	0.3707	0.8945

MSE 비교 결과, XGboost 모델이 가장 낮은 MSE로 적합한 모델로 판단되었다.

Part 3 결론 및 시사점

Driven Data에서 제공한 뎅기열 확산건수 데이터를 활용하여 온도, 강수량, 초목 등의 환경 변수에 따라 뎅기열 확산수를 예측할 수 있었다. XGboost가 가장 적합한 모델로 판단되었고, 대부분 확산수에 영향을 미치는 변수는 시계열성을 나타내는 년도, 주와 온도에 관한 변수였다.

뎅기열 확산 수를 예측하므로써 전세계의 생명을 위협하는 전염병에 대처할 수 있을 것이며, 더 나아가 기후 환경 변화에 대한 경각심 또한 불러일으킬 수 있을 것이다.

베이지안 분위 회귀 모형과 ML 방법을 도입하였으나, 생각보다 적합하지 않아 아쉬움이 남는다. 하지만, tuning에 대하여 조금만 더 신경을 쓴다면, 더 적합한 모델을 찾을 수 있을 것이라 생각된다.