팀 주제: 일교차 예측에 따른 감기 예방

수업명: 빅데이터 시계열 모형

데이터사이언스학부 김필준, 데이터사이언스학부 심창우, 데이터사이언스학부 김성우, 소프트웨어학부 박현, 데이터사이언스학부 강민지

지도교수 : 박현숙

요약

기상청 일별 데이터와 국민건강보험공단의 데이터를 이용하여 TBATS 모형을 사용하여 일교차를 예측했고, 감기와 환절기의 상관관계를 구하였다.

연 구 배경

- 연구배경: 실제로 코로나19 대유행 시기를 보면 대유행 발생시기가 환절기에 나타난 것을 볼 수 있다. 환절기에는 여러 호흡기 질환에 취약하다. 따라서, 환절기 시기에 늘어난 감기 환자의 추세를 파악하고, 예측을 통해 여러 질병을 예방하고자 주제를 선정했다.
- 연구주제: 일교차 예측에 따른 감기 예방

연 구 방 법

데이터 설명

기상청: 춘천에 2015년 1월 1일부터 2019년 12월 31일까지의 일별기온과 풍속 등 기존의 변수들을 이용하여 체감온도로 칼럼이 구성되어 있다. 12월 31일은 행의 개수가 하나이기 때문에 일교차를 확인할 수 없어 제거했다.
국민건강보험공단: 춘천시 일별 감기 환자 데이터.

	d	t	c	w	s
	<chr>	<dttm>	<dbl>	<dbl>	<dbl>
1	2015-01-01	2015-01-01 01:00:00	-6.8	3.5	-8.294292
2	2015-01-01	2015-01-01 02:00:00	-7.5	3.2	-8.818981
3	2015-01-01	2015-01-01 03:00:00	-8.0	4.0	-10.005223
4	2015-01-01	2015-01-01 04:00:00	-8.5	2.7	-9.441927
5	2015-01-01	2015-01-01 05:00:00	-9.0	1.8	-8.885087
6	2015-01-01	2015-01-01 06:00:00	-10.0	0.6	-7.226491

	date	name	count	value
	<int>	<chr>	<int>	<dbl>
1	20150101	춘천시	389	6.713558
2	20150102	춘천시	2886	16.653787
3	20150103	춘천시	1761	16.372755
4	20150104	춘천시	411	7.264668
5	20150105	춘천시	2572	10.906344
6	20150106	춘천시	1850	12.226551

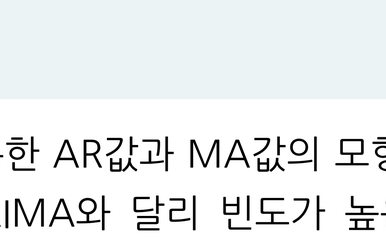
분석방법

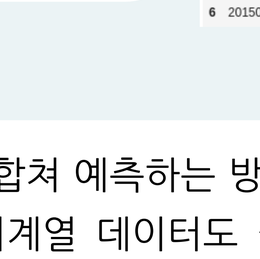
차분을 통한 AR값과 MA값의 모형을 합쳐 예측하는 방법인 ARIMA가 있고, ARIMA와 달리 빈도가 높은 시계열 데이터도 설명할 수 있는 TBATS가 있다.

TBATS 사용 이유: ARIMA는 Frequency를 360까지 지원을 하기 때문에 1년이 365일인 일별 데이터를 분석하기에 부적합하다 생각이 되어 Frequency가 365를 넘어서도 지원을 하는 TBATS를 사용했다.

연 구 결 과

결과 1

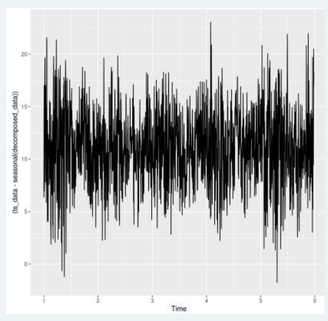


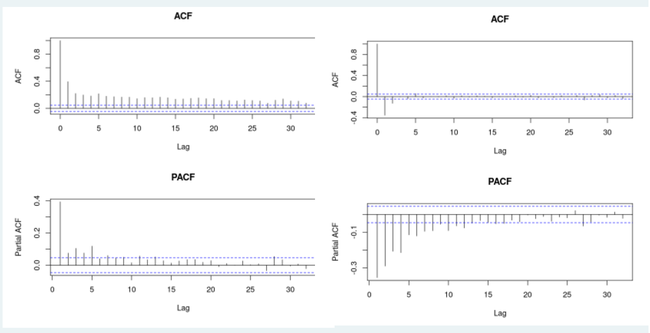


1. 일교차 데이터를 가지고 쿼리를 추출한 시각화이다. 1년 주기로 계절성이 나타난다는 사실을 볼 수 있다.
2. 시계열을 계절성과 나머지 파트로 나눠서 분해했다.

연 구 결 과

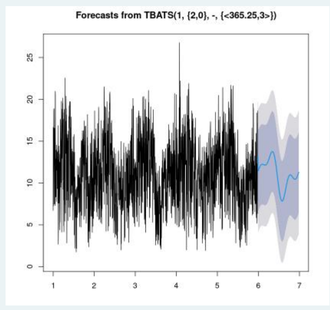
결과 2





1. 분해했던 데이터 중 계절성을 추출하여 본래의 데이터에서 차감을 해 시각화 한 결과이다.
2. 자기상관분석이므로, ARIMA 모형의 q와 p를 결정할 수 있다. 한 번 차분을 한 결과, acf의 lag가 1 이후로부터 정상성을 보이는 것을 알 수 있고, 차분 전에는 pacf의 lag가 5 이후일 때부터 정상성을 이룬다.

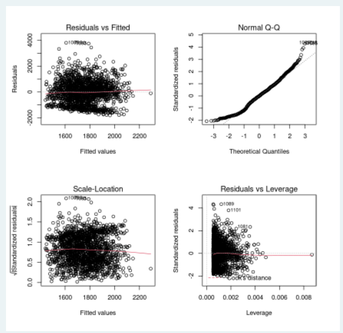
결과 3




```
cor.test(fdf$count, fdf$value)

        Pearson's product-moment correlation

data:  fdf$count and fdf$value
t = 6.7124, df = 1819, p-value = 2.55e-11
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.1103237 0.1999780
sample estimates:
      cor
0.155471
```



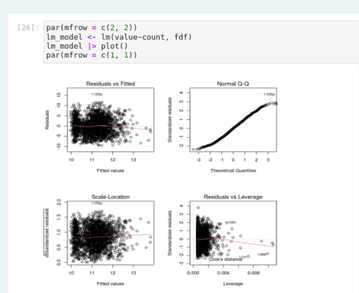
1. 주기가 1년이라는 기간때문에 ARIMA 모형을 사용하여 시계열 예측을 하는 것에 무리가 있었다. 그래서, Frequency가 365를 넘어도 지원을 하는 TBATS라는 모형을 사용한 결과이다. 시계열 모형을 이용하여 데이터를 예측했다.
2. 감기와 일교차에 대한 상관분석이다. 상관계수 중에서 대표적인 피어슨 상관계수를 활용한 결과 P-value 값이 0.15로 낮게 나왔다.
3. 선형 회귀 모형을 이용하여 일교차와 감기의 진찰 수에 대한 회귀분석을 했다.

논 의 및 결 론

감기 환자의 예측은 할 수 있으나, 상관계수가 낮아 도입하기에는 어려움이 있다.

하지만 상관계수가 작은 편에 속함에도 어느 정도의 의미있는 상관성이 있다고 보여지기 때문에 일교차가 심한 환절기에는 조심해야 한다

```
[26]: par(mfrow = c(2, 2))
      lm_model <- lm(value~count, fdf)
      lm_model |> plot()
      par(mfrow = c(1, 1))
```




```
[27]: lm_model |> summary()

Call:
lm(formula = value ~ count, data = fdf)

Residuals:
     Min       1Q   Median       3Q      Max
-9.4716 -2.9797  0.0053  2.8230 15.3332

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 9.8776023  0.2089849  47.265  < 2e-16 ***
count       0.0007115  0.0001060   6.712 2.55e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.019 on 1819 degrees of freedom
Multiple R-squared:  0.02417,   Adjusted R-squared:  0.02363
F-statistic: 45.06 on 1 and 1819 DF,  p-value: 2.55e-11
```