최종작성: 21. 11. 14(일)

작성자	검토자

사례연구 3

State.77 데이터셋 다중 회귀분석
및
코스피지수를 활용한 시계열 분석

B2 조: 문현진(조장) 남원식 오준서

21. 11. 15(월) 제출

목 차

서론

R 기본 내장 데이터 셋인 State.x77 과 한국거래소에서 제공하는 과거 10 년간의 코스피 지수 데이터 셋을 활용하여 각각 다중 회귀분석과 시계열 분석을 시행한 보고서이다.

1) State data sets

state 데이터 셋에는 state.abb, ,area, center, division, name, region, x77 데이터 셋이 존재한다. 그 중 state.x77데이터 셋을 이용하여 다중회귀분석을 실시하고 유의미한 독립변수를 걸러내어 기대수명에 가장 큰 영향을 미치는 요인들을 살펴본다.

2) KOSPI 데이터

한국거래소에서 제공하는 많은 데이터 중 과거 10년(2011.11.08~2021.11.08)간의 코스피 지수 데이터를 가져온다. 이 자료는 비정상성 시계열 자료로서 시계열 분석을 수행하여 추세와 계절성 요인 등을 파악하고 미래를 예측한다.

본론

1. State.x77 데이터 셋 다중회귀분석

1) 데이터 셋 준비

State.x77 데이터 셋 내용

	Population	Income	Illiteracy	Life.Exp	Murder	HS.Grad	Frost	Area
Alabama	3615	3624	2.1	69.05	15.1	41.3	20	50708
Alaska	365	6315	1.5	69.31	11.3	66.7	152	566432
Arizona	2212	4530	1.8	70.55	7.8	58.1	15	113417
Arkansas	2110	3378	1.9	70.66	10.1	39.9	65	51945
California	21198	5114	1.1	71.71	10.3	62.6	20	156361
Colorado	2541	4884	0.7	72.06	6.8	63.9	166	103766
Connecticut	3100	5348	1.1	72.48	3.1	56	139	4862
Delaware	579	4809	0.9	70.06	6.2	54.6	103	1982
Florida	8277	4815	1.3	70.66	10.7	52.6	11	54090
Georgia	4931	4091	2	68.54	13.9	40.6	60	58073

(중략)

Washington	3559	4864	0.6	71.72	4.3	63.5	32	66570
West Virginia	1799	3617	1.4	69.48	6.7	41.6	100	24070
Wisconsin	4589	4468	0.7	72.48	3	54.5	149	54464
Wyoming	376	4566	0.6	70.29	6.9	62.9	173	97203

(표1-1)

USA 의 각 state 에 대한 Population, Income, Illiteracy, Life,Exp, murder, HS.Grad, Frost, Area 에 대한 정보를 확인한다.

2) 기대수명 변수에 대한 회귀분석(1)

Life.Exp를 **종속변수**로 설정하고 **나머지 모든 변수**를 **독립변수**로 넣어 회귀분석을 실시한 결과이다 .

```
Call:
lm(formula = model_1, data = state)

Residuals:
     Min       1Q   Median       3Q      Max
-1.48895 -0.51232 -0.02747  0.57002  1.49447

(표 1-2)

Residual standard error: 0.7448 on 42 degrees of freedom
Multiple R-squared:  0.7362,  Adjusted R-squared:  0.6922
F-statistic: 16.74 on 7 and 42 DF,  p-value: 2.534e-10
```

이 모델에서 F-통계량=16.74, p-value=2.534e-10 이므로

Life.Exp에 대한 독립변수들 간의 모형은 유의수준 5%하에서 통계적으로 매우 유의하다.

Summary	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	7.094e+01	1.748e+00	40.586	< 2e-16
Population	5.180e-05	2.919e-05	1.775	0.0832
Income	-2.180e-05	2.444e-04	-0.089	0.9293
Illiteracy	3.382e-02	3.663e-01	0.092	0.9269
Murder	-3.011e-01	4.662e-02	-6.459	8.68e-08
HS.Grad	4.893e-02	2.332e-02	2.098	0.0420
Frost	-5.735e-03	3.143e-03	-1.825.	0.0752
Area	-7.383e-08	1.668e-06	-0.044	0.9649

(표 1-2)

결과)

회귀분석 결과로 추정된 회귀식은

[기대수명 = 70.94 + 0.00005180(Population) - 0.0002180(Income) + 0.03382(Illiteracy) - 0.3011(Murder) +0.04893(HS.Grad) – 0.005735(Frost) – 0.00000007383(Area)]이다.

따라서,

기대수명을 **증가**시키는 요인으로는 Population, Illiteracy, HS.Grad가 있고

기대수명을 **감소**시키는 요인으로는 Income, Murder, Frost, Area가 있음을 알 수 있다.

그러나, 회귀계수의 p-value값이 0.05보다 큰 변수들이 존재하는 것으로 보아 모든 변수가 통계적으로 유의하지는 않다는 것을 확인할 수 있다.

(1)번 회귀모델에 대한 시각화 및 설명)

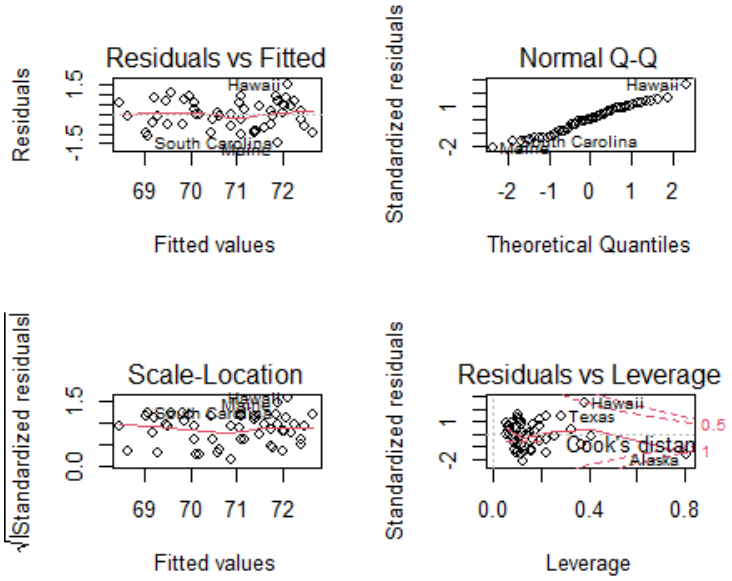


그림1-1)

각 분석결과에 따른 해석은 아래와 같다.

(1) Residuals vs Fitted

x축은 회귀모형을 통해 예측된 값(기대수명)이고, y축은 잔차이다. 선형회귀 모형은 오차가 정규분포를 따른다는 정규성을 가정하므로 기울기가 0에 가까운 이 그래프는 이상적인 모형에 가깝다고 볼 수 있다.

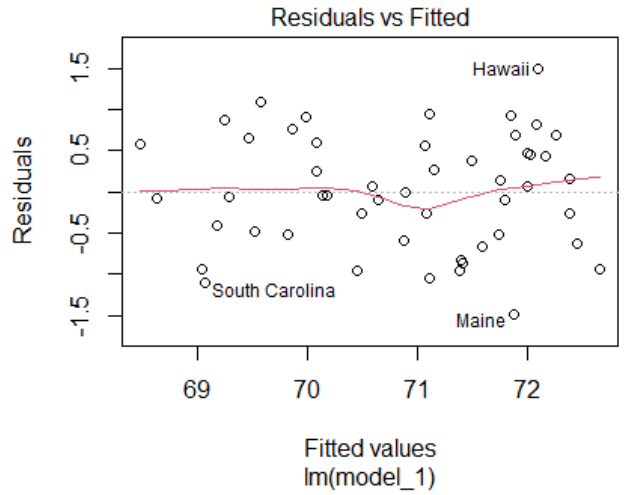


그림1-2)

(2) Normal Q-Q

표준화된 잔차의 확률도를 나타내는 그래프의 점들이 45도 각도의 직선을 이루는 형태를 띄므로 이 모델은 정규성 가정을 만족한다고 볼 수 있다.

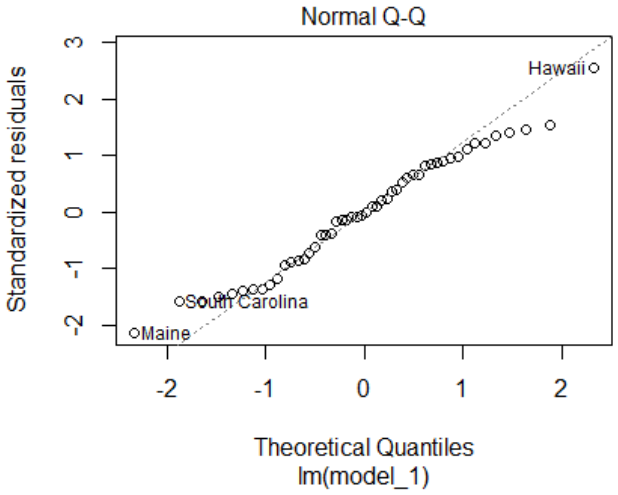


그림1-3)

(3) Scale-Location

x축은 회귀모형을 통해 예측된 y값이며, y축은 표준화 잔차를 나타낸다. 기울기가 0인 직선의 형태가 관측되는 것은 이상적이다. 그러나 Hawaii 및 소수의 점들이 매우 멀리 떨어져 있는데 이 지점에서는 회귀모형이 예측을 잘 하지 못했음을 알 수 있다.

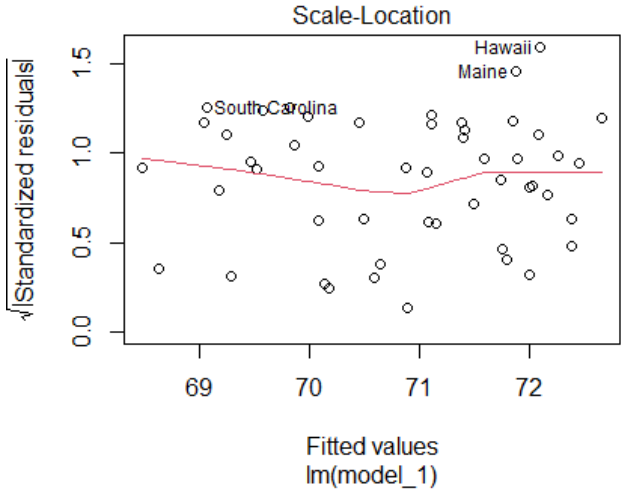


그림1-4)

(4) Residuals vs Leverage

x축은 레버리지, y축은 표준화 잔차값을 나타낸다. 레버리지란 관측치가 다른 관측치 집단으로부터 떨어진 정도를 나타내며 독립변수가 얼마나 극단에 치우쳐 있는지를 보여준다. 여기서 0.5 이상인 빨간 점선의 밖에 있는 Hawaii와 Alaska는 예측치를 크게 벗어난 관측치이다.

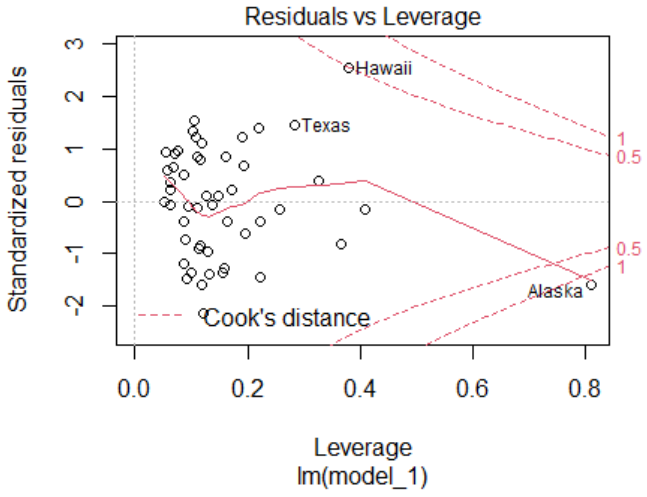


그림 1-5)

정리)

(1)번 모델은 정규성 가정을 만족하지만 무의미한 변수가 섞여있고, Leverage 를 나타낸 4 번째 그래프에서 예측치를 크게 벗어난 관측치들이 발견되는 것으로 보아 해당 모델은 아직 보완이 필요한 것을 알 수 있다.

다음 장에서는 p-value 값이 0.5 이상인 변수를 제외하고 회귀모델을 만들어 분석을 실시한다.

3) 기대수명 변수에 대한 회귀분석(2)

Life.Exp를 **종속변수**로 설정하고 (1)번 모델에서 p-value값이 0.5 이상인 **Income, Illiteracy, Area를 제외한 나머지**를 **독립변수**로 넣어 회귀분석을 실시한 결과이다 .

```
Call:
lm(formula = for2, data = state)

Residuals:
     Min       1Q   Median       3Q      Max
-1.47095 -0.53464 -0.03701  0.57621  1.50683

(표 1-3)

Residual standard error: 0.7197 on 45 degrees of freedom
Multiple R-squared:  0.736,    Adjusted R-squared:  0.7126
F-statistic: 31.37 on 4 and 45 DF,  p-value: 1.696e-12
```

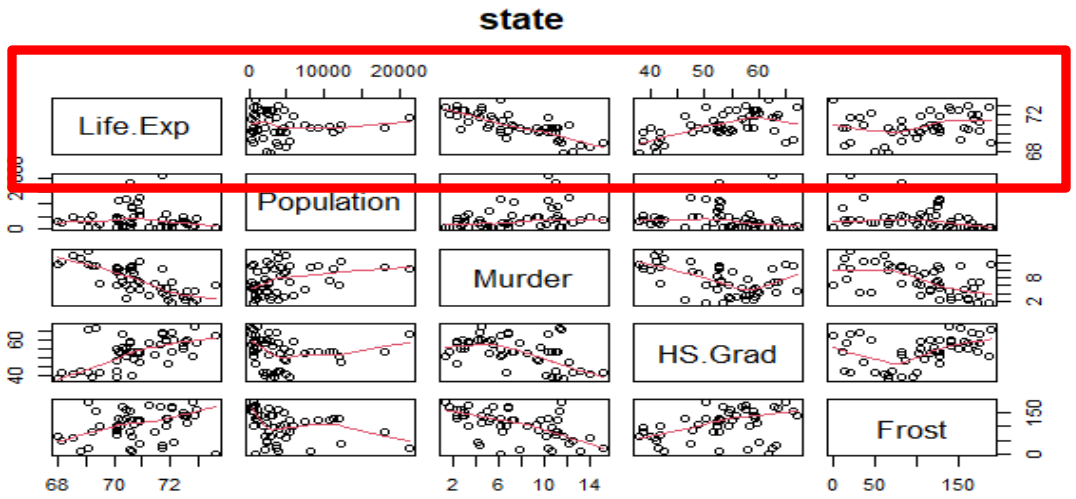
이 모델에서 F-통계량=31.37, p-value=1.696e-12 이므로

Life.Exp에 대한 독립변수들 간의 모형은 유의수준 5%하에서 통계적으로 매우 유의하다.

Summary	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	7.103e+01	9.529e-01	74.542	< 2e-16
Population	5.014e-05	2.512e-05	1.996	0.05201
Murder	-3.001e-01	3.661e-02	-8.199	1.77e-10
HS.Grad	4.658e-02	1.483e-02	3.142	0.00297
Frost	-5.943e-03	2.421e-03	-2.455.	0.01802

(표 1-3)

각 변수들 간의 산점도



(그림 1-6)

결과)

회귀분석 결과로 추정된 회귀식은

[기대수명 = 70.103 + 0.00005014(Population) - 0.3001(Murder) +0.04893(HS.Grad)]이다.

따라서,

기대수명을 **증가**시키는 요인으로는 Population, HS.Grad가 있고

기대수명을 **감소**시키는 요인으로는 Murder, Frost가 있음을 알 수 있다.

(2)번 모델은 (1)번 모델보다 모형의 설명력이 69.22% > 71.26%로 2.04%p 증가했다.

+ 회귀분석 전 (그림1-6)의 산점도를 살펴보면 상대적으로 관련성이 적은 변수들이 있음을 대략적으로 알 수 있다.

정리)

(1)번모델에서 무의미한 변수들을 제거하고 (2)번모델을 생성하여 분석하였다.

다음 장에서는 (2)번 모델에서도 관련도가 가장 높은 2 개의 변수로 회귀모델을 만들어 분석을 실시한다.

4) 기대수명 변수에 대한 회귀분석(3)

Life.Exp를 **종속변수**로 설정하고 (2)번 모델에서 관련성이 가장 높은

Murder와 **HS.Grad변수**를 **독립변수**로 넣어 회귀분석을 실시한 결과이다 .

```
Call:
lm(formula = for3, data = state)

Residuals:
     Min       1Q   Median       3Q      Max
-1.66758 -0.41801  0.05602  0.55913  2.05625

(표 1-4)

Residual standard error: 0.7959 on 47 degrees of freedom
Multiple R-squared:  0.6628,  Adjusted R-squared:  0.6485
F-statistic:  46.2 on 2 and 47 DF,  p-value: 8.016e-12
```

이 모델에서 F-통계량=46.2, p-value=8.016e-12 이므로

Life.Exp에 대한 독립변수들 간의 모형은 유의수준 5%하에서 통계적으로 매우 유의하다.

Summary	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	70.29708	1.01567	69.213	< 2e-16
Murder	-0.23709	0.03529	-6.719	2.18e-08
HS.Grad	0.04389	0.01613	2.721	0.00909

(표 1-4)

결과)

회귀분석 결과로 추정된 회귀식은

[기대수명 = 70.29708 - 0.23709(Murder) +0.04389(HS.Grad)]이다.

따라서,

기대수명에 **가장 큰 영향을 미치는** 요인은 **Murder**이고,

살인비율이 10만명당 1명일 때 **기대수명**은 *0.23709*살 **감소**하고,

고졸비율이 1% 증가할 때마다 **기대수명**이 *0.04389*살 **증가**하는 것을 알 수 있다.

5) 기대수명의 예측 결과값

예시) 전 인구의 55%가 고졸이고 살인비율이 10만명당 8명일 때 기대수명의 결과값은?

기대수명=70.29708+0.04389(HS.grad)-0.23709(murder)이므로

직접 계산했을 때,

각각 고졸비율에 55, 살인비율에 8을 넣으면

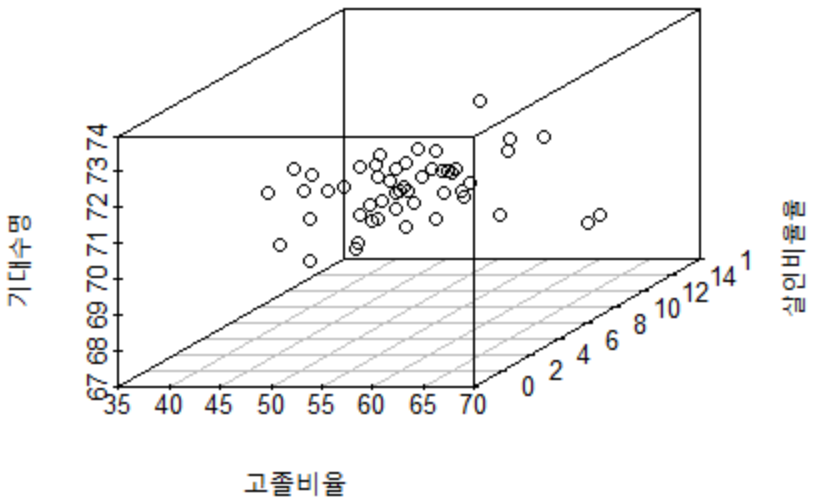
기대수명=70.29708+2.41395-1.89672

= 70.81431살 이라는 예측값을 얻을 수 있다.

또는,

Predict()함수를 통해 구하면 70.81416살 이라는 예측값을 얻을 수 있다.(부록 코드1-5)

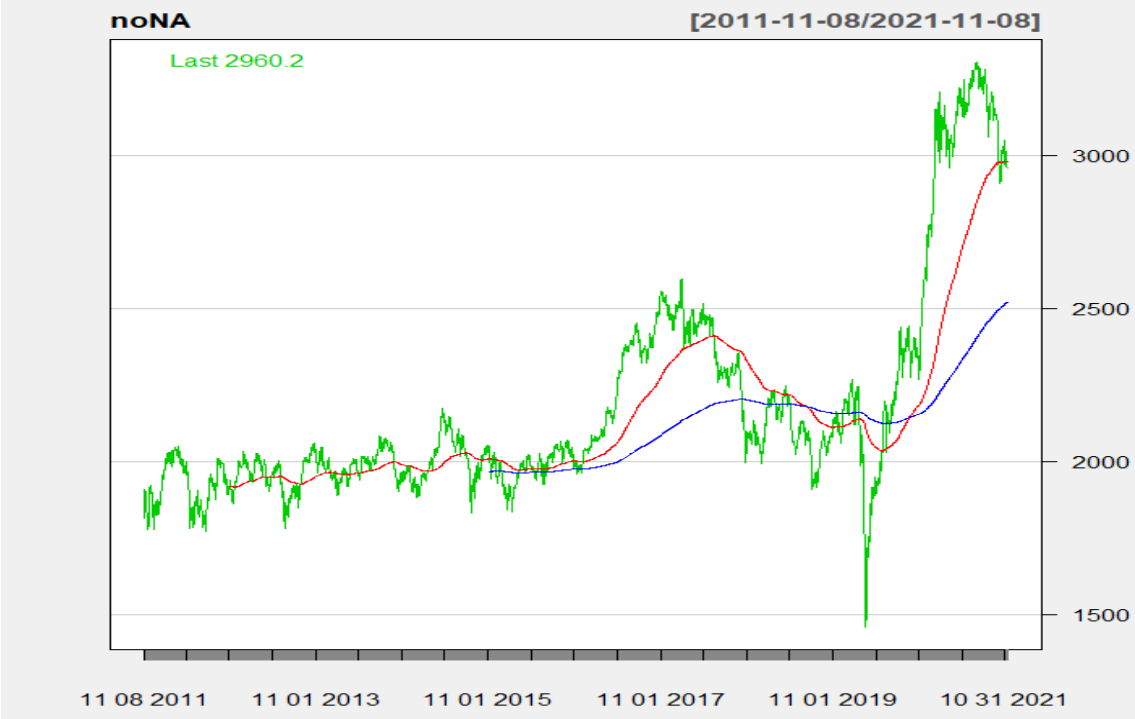
6) 회귀모델 3D 그래프 시각화



(그림1-7)

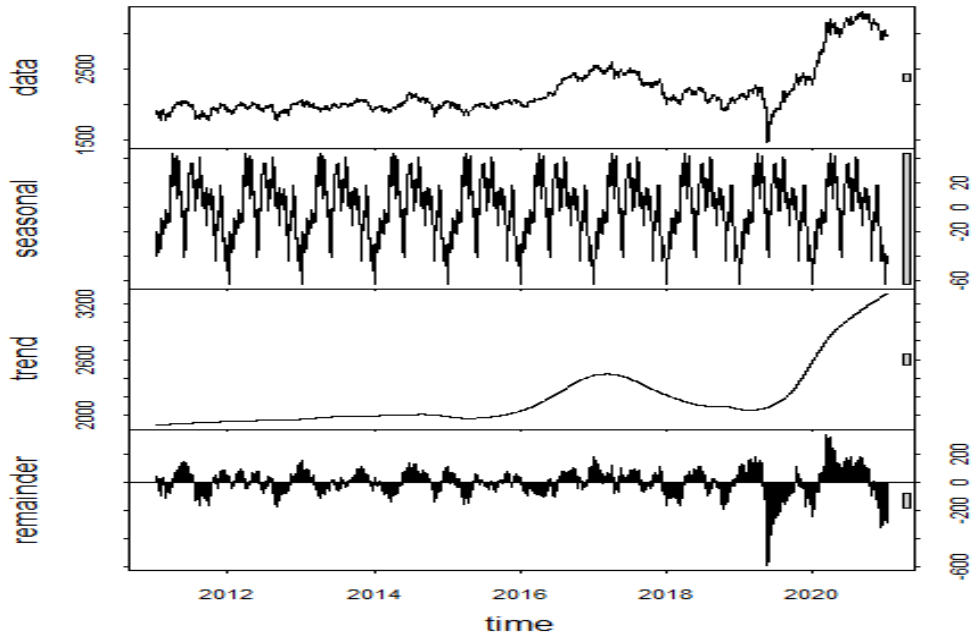
2. KOSPI지수를 활용한 시계열분석

1) 추세선 확인



(그림2-1)

2) 시계열 자료 변동요인 분해



(그림2-2)

3) 결과 해석

빨간선은 1년단위 추세선이고, 파란선은 4년단위 추세선을 그린 결과이다. 2017년에 큰 폭으로 상승한뒤 평균으로 회귀하는 모습을 보이다가 19년도를 전후로 큰 등락을 겪었다.

최근 지수는 점점 하락하는 모습을 보이고 있지만 연단위 추세선에서는 아직 상승하는 모습이다. 시계열 자료의 변동요인을 분해한 (그림2-2)를 보면 data는 KOSPI지수의 관측값들을 나타내고 있고, seasonal 그래프를 보면 계절성분을 갖고있다고 보기 어려운 그래프 형태를 띄고있다.

Trend 그래프에서는 초기에 완만한 증가세를 보이다가 중반부터는 급격한 증가 및 감소가 이어지고 말기에는 급격한 증가세가 눈에 띈다.

Remainder 그래프에서는 불규칙 성분인 잔차를 확인할 수 있다.

결론

R 에서 기본적으로 제공하는 state.x77 데이터 셋과 한국거래소에서 제공하는 코스피 지수 데이터를 각각 회귀분석과 시계열 분석을 실시하였다.

state.x77 데이터 셋을 분석한 결과 살인비율과 고졸비율이 기대수명에 가장 큰 영향을 미치는 변수임을 파악하고 각 독립변수의 변화에 따른 기대수명의 예측치까지 구할 수 있었다.

코스피 데이터의 시계열 분석에서는 이산적으로 분리된 시계열 데이터를 연속적으로 만들기 위해 주말 및 공휴일의 결측치를 예상치로 대체하여 분석을 진행하였으나, 관측시점 사이의 간격이 너무 촘촘하여 ARIMA 분석을 수행할 수 없었던 것이 아쉬웠다.

부록

전체코드

```
## B2 조 사례연구 3 ####

# study3.2.0
# 변경내용
  # 결측치 오류처리 해결
  # 변수명 일부 변경


rm(list=ls())
getwd()
setwd('C:/rwork/')
install.packages("car")
install.packages('scatterplot3d')
install.packages("tidyverse")
install.packages("dplyr")
install.packages("zoo")
install.packages("forecast")
library(scatterplot3d)
library(dplyr)
library(car)
library(tidyverse)
library(zoo)
library(forecast)
library(quantmod)
```

```
## 문제 1 ##############################################################################
# 다음 사항을 적용하여 다중회귀분석을 실시하시오

# 1) state 데이터 셋을 load 하고, state.77 data.set 을 데이터프레임으로 변환하고,
# life EXP 변수를 Life.EXP 로 HS Grad 변수를 HS.Grad 로 변경하시오.
state <- data.frame(state.x77)
str(state)


# 2) Life Expectancy 변수를 종속변수로 설정하고 나머지 변수를 독립변수로 설정하여
# 회귀분석을 실시하시오. 실시 후 결과에 대해 해석하시오
names(state)
model_1 = Life.Exp ~ Population+Income+Illiteracy+Murder+HS.Grad+Frost+Area
state_1.lm <- lm(formula = model_1, data=state)  # (1)번모델 생성
```

```
summary(state_1.lm)  # 회귀모델의 요약정보
vif(state_1.lm) # 공선성 확인
cor(state) # 상관분석결과
par(mfrow=c(2,2))
plot(state_1.lm)  # 회귀분석 그래프 시각화


# 3) 2)번 회귀모형에서 Income, Illiteracy, Area 변수를 제외하고 회귀분석을 실시하고
# 결과에 대해 해석하시오
for2 = Life.Exp ~ Population+Murder+HS.Grad+Frost
state_2.lm <- lm(formula = for2, data=state)  # (2)번모델 생성
summary(state_2.lm)
  # 각 변수간의 산점도 출력
newstate2 <- state %>% select(1,4,5,6,7) %>% pairs(panel=panel.smooth, main='state')


# 4) Life Expectancy 변수를 종속변수로 설정하고 HS.Grad 와 Murder 변수를 예측변수
# (predictor variable)로 설정하여 회귀분석을 실시하시오
for3 = Life.Exp ~ Murder+HS.Grad
state_3.lm <- lm(formula = for3, data=state)  # (3)번모델 생성
plot(state_3.lm)
summary(state.lm)


# 5) 전 인구의 55%가 고졸이고 살인비율이 10 만명당 8 명일 때 Life Expectancy 결과값을
# 예측하시오
summary(state_3.lm)
# 기대수명=70.29708+0.04389(HS.grad)-0.23709(murder)이므로
# 각각 고졸비율에 55, 살인비율에 8 을 넣으면
# life=70.29708+2.41395-1.89672
#     =70.81431 살 이라는 예측값을 얻을 수 있다.
newpred <- predict(state_3.lm, newdata=data.frame(HS.Grad=55,Murder=8))
newpred  # 기대수명=70.81416 살

# 6) 4)번에서처럼 2 개의 독립변수, 1 개의 종속변수의 데이터와 fit 된 회귀평면을
# 3D 그래프로 시각화 하시오.
scatterplot3d(state$Life.Exp~state$HS.Grad+state$Murder, highlight.3d = FALSE,
              type = "p",xlab='고졸비율',ylab = '살인비율',zlab = '기대수명')
```

```
## 문제 2 ###############################################################################
# 2-1) 한국거래소에서 받아온 코스피 데이터를 변수에 저장한다.
kospi <- read.csv('kospi_data_10year.csv', header=T)


# 2-2) 코스피 데이터 셋에서 일자와 종가 데이터를 ko.date 변수에 저장한다.
ko.date <-kospi %>% select(1,2) %>% as.data.frame
ko.date$일자 <- as.Date(ko.date$일자)


# 2-3) 빠져있는 주말과 공휴일을 생성하여 병합한다.
calender <- seq.Date(min(ko.date$일자),max(ko.date$일자),"day") %>%
  data.frame() %>% `colnames<-`('일자')

day365 <- full_join(calender, ko.date, by='일자')


# 2-4) 결측치를 유사값으로 대체한다.
zooval <- zoo(day365$종가,day365$일자)
noNA <- zooval %>% na.approx %>% data.frame


# 2-5) 그래프를 생성하여 추세선을 확인한다.
chartSeries(noNA, theme=chartTheme('white'),
            type = c('auto', 'mathsticks'),
            subset = '2011-11::',
            show.grid=TRUE,
            major.ticks = 'auto',minor.ticks = TRUE,
            multi.col = F,
            TA="addEMA(365.25,col='red');addEMA(1461,col='blue')")


# 2-6) 시계열 데이터를 생성한다.
result.ts<- ts(data = noNA,
               start = c(2011,11,08),end = c(2021,11,08), frequency =365)


# 2-7) 시계열 요소 분해 및 해석
# 기본/ 계절변동/ 추세변동/ 잔차
plot(stl(result.ts,'periodic'))
```

표 1-1 전체 데이터 셋

	Population	Income	Illiteracy	Life.Exp	Murder	HS.Grad	Frost	Area
Alabama	3615	3624	2.1	69.05	15.1	41.3	20	50708
Alaska	365	6315	1.5	69.31	11.3	66.7	152	566432
Arizona	2212	4530	1.8	70.55	7.8	58.1	15	113417
Arkansas	2110	3378	1.9	70.66	10.1	39.9	65	51945
California	21198	5114	1.1	71.71	10.3	62.6	20	156361
Colorado	2541	4884	0.7	72.06	6.8	63.9	166	103766
Connecticut	3100	5348	1.1	72.48	3.1	56	139	4862
Delaware	579	4809	0.9	70.06	6.2	54.6	103	1982
Florida	8277	4815	1.3	70.66	10.7	52.6	11	54090
Georgia	4931	4091	2	68.54	13.9	40.6	60	58073
Hawaii	868	4963	1.9	73.6	6.2	61.9	0	6425
Idaho	813	4119	0.6	71.87	5.3	59.5	126	82677
Illinois	11197	5107	0.9	70.14	10.3	52.6	127	55748
Indiana	5313	4458	0.7	70.88	7.1	52.9	122	36097
Iowa	2861	4628	0.5	72.56	2.3	59	140	55941
Kansas	2280	4669	0.6	72.58	4.5	59.9	114	81787
Kentucky	3387	3712	1.6	70.1	10.6	38.5	95	39650
Louisiana	3806	3545	2.8	68.76	13.2	42.2	12	44930
Maine	1058	3694	0.7	70.39	2.7	54.7	161	30920
Maryland	4122	5299	0.9	70.22	8.5	52.3	101	9891
Massachusetts	5814	4755	1.1	71.83	3.3	58.5	103	7826
Michigan	9111	4751	0.9	70.63	11.1	52.8	125	56817
Minnesota	3921	4675	0.6	72.96	2.3	57.6	160	79289
Mississippi	2341	3098	2.4	68.09	12.5	41	50	47296
Missouri	4767	4254	0.8	70.69	9.3	48.8	108	68995
Montana	746	4347	0.6	70.56	5	59.2	155	145587
Nebraska	1544	4508	0.6	72.6	2.9	59.3	139	76483
Nevada	590	5149	0.5	69.03	11.5	65.2	188	109889
New Hampshire	812	4281	0.7	71.23	3.3	57.6	174	9027
New Jersey	7333	5237	1.1	70.93	5.2	52.5	115	7521
New Mexico	1144	3601	2.2	70.32	9.7	55.2	120	121412
New York	18076	4903	1.4	70.55	10.9	52.7	82	47831
North Carolina	5441	3875	1.8	69.21	11.1	38.5	80	48798
North Dakota	637	5087	0.8	72.78	1.4	50.3	186	69273
Ohio	10735	4561	0.8	70.82	7.4	53.2	124	40975
Oklahoma	2715	3983	1.1	71.42	6.4	51.6	82	68782

Oregon	2284	4660	0.6	72.13	4.2	60	44	96184
Pennsylvania	11860	4449	1	70.43	6.1	50.2	126	44966
Rhode Island	931	4558	1.3	71.9	2.4	46.4	127	1049
South Carolina	2816	3635	2.3	67.96	11.6	37.8	65	30225
South Dakota	681	4167	0.5	72.08	1.7	53.3	172	75955
Tennessee	4173	3821	1.7	70.11	11	41.8	70	41328
Texas	12237	4188	2.2	70.9	12.2	47.4	35	262134
Utah	1203	4022	0.6	72.9	4.5	67.3	137	82096
Vermont	472	3907	0.6	71.64	5.5	57.1	168	9267
Virginia	4981	4701	1.4	70.08	9.5	47.8	85	39780
Washington	3559	4864	0.6	71.72	4.3	63.5	32	66570
West Virginia	1799	3617	1.4	69.48	6.7	41.6	100	24070
Wisconsin	4589	4468	0.7	72.48	3	54.5	149	54464
Wyoming	376	4566	0.6	70.29	6.9	62.9	173	97203

```
## test zone ########################################################################
# 1-1
# state.x77 데이터 셋 확인 과정
state <- data.frame(state.x77)
write.csv(state,'state.csv',quote = F)

# 1-2
plot(state_1.lm,which = c(1:6))  # Cock's distance 관련 정보까지 출력

# 1-4
newstate <- state %>% select(4,5,6) %>% pairs(panel=panel.smooth, main='state')
# 산점도를 그려 확인해보면 살인률과 고졸비율은 기대수명과는 밀접한 관련이 있지만
# 서로에게는 관련성이 적은 것을 확인할 수 있다.

pred<- predict(state_3.lm, state)
pred  # (3)번모델의 각 주별 기대수명 예측치

# 1-7
plot(decompose(result.ts))  # stl()과 유사한 방법
```

참고자료

1) 결측치 대체
https://stackoverflow.com/questions/27368195/r-ts-with-missing-values

2) 코스피 데이터 셋
http://www.krx.co.kr/main/main.jsp