[예측방법론]

제목 : 비경제활동인구를 이용한 분석과 예측을 통한 인사이트 제시

응용통계학과
32142221-서근태

1. 비경제활동인구의 분석결과 요약(summary)

비경제활동인구의 모형 제작을 할 때, 모형식별과정에서 정상성과 계절성은 과대차분을 피해 차분 1을 선택, 모수 q, P는 모수 간결성을 위해 1로 선택하여 sarima모형이 결정되었습니다.

이후 sarima모형에서의 모수(계수)를 추정하였고 진단과정에서 과대적합이 일어나지 않아 모형을 확신할 수 있었습니다. 마지막으로 잔차분석에서 데이터가 자기상관관계를 갖지 못하고 상관계수가 0임을 알 수 있었습니다. 결과적으로 최종모형으로 sarima(0,1,1)(1.1.0)(12)인 $(1-0.37063B^{12})\ln Z_t = (1-0.22614B)\ln\epsilon_t$을 선택했습니다.

이 모형을 통해 2020년 10월부터 2022년 5월까지를 예측한다면, 불규칙성분인 코로나로 인해 비경제활동인구가 증가하는 모습을 볼 수 있습니다. 마지막으로 저는 시계열 분석을 톻해서 예측은 항상 정답이 아닌 미래결과의 보기 중 하나라는 것을 명심할 수 있었습니다.

2. 비경제활동인구의 내용 분석

비경제 활동인구 : 만 15세 이상 총인구 중에서 취업자도 실업자도 아닌 사람, 즉 일할 능력이 있어도 일할 의사가 없거나, 일할 능력이 없는 사람들을 말한다. [위키백과]

하지만 저는 이를 사용할 때 비경제활동인구는 취업준비자들이 몇 번의 취업실패를 겪으면서 취업에 대한 마음의 변화가 생기는 분을 중심으로 생각했습니다.

비경제활동인구의 특징은 대규모 채용이 일어나는 3월~ 11월 사이에 적은 인구수를 보이는 것입니다. 보통 전반기 채용(3월), 하반기 채용(9월)이 시작되므로 그 기간 큰 변화없이 어느정도 평활한 모습을 보여줍니다. 3월부터 채용이 시작되어 확정이 되는 6월에 가장 많은 취업인구수를 기록함과(경제활동인구수 참조) 동시에 비경제활동인구는 가장 낮은 인구수를 기록했습니다.

이와 반대로 12월 ~2월 사이에 가장 많은 비경제활동인구 수를 보여줍니다. 이는 보통 연말에 행해지는 정년은퇴나 일년을 마무리하는데 필요한 단기간의 인턴/일용직 노동자의 해고와 관련되어 있다고 생각했습니다. 또한 채용시장이 얼어붙은 시기인지라 그만큼 취업의 사가 떨어져 비경제활동인구 수가 급상승 하는 그래프를 보여줍니다.

하지만 비경제활동인구수 그래프에서 기존과 다르게 큰 변화가 있는 2개의 연도가 있었습니다. 먼저 2013년, 비경제활동인구수가 가장 높은 수치를 기록하게 되는데 이는 글로벌 금융위기가 터지고 중국회사들과의 가격경쟁에서 패배하면서 많은 기업이 사라진 것이 원인이

었습니다. 이는 곧 취직 기회가 줄어들었고 취직에 대한 불안감이 일할 의사와 연관되어 비경제활동인구수 증가에 큰 영향을 끼친 것으로 봤습니다.

이에 대기업과 공공기관은 채용규모를 소폭 늘린 것으로 밝혀졌습니다. 하지만 대졸 정규직 이외에도 고졸, 시간선택제 등 종류가 다양해지고, 주요 그룹 이외의 그룹사는 채용인원을 줄였습니다. 이는 체감되는 구직시장 개선이 크지 않아 비경제활동인구수를 줄이기 위한 방안으로 적절하지 않았습니다.

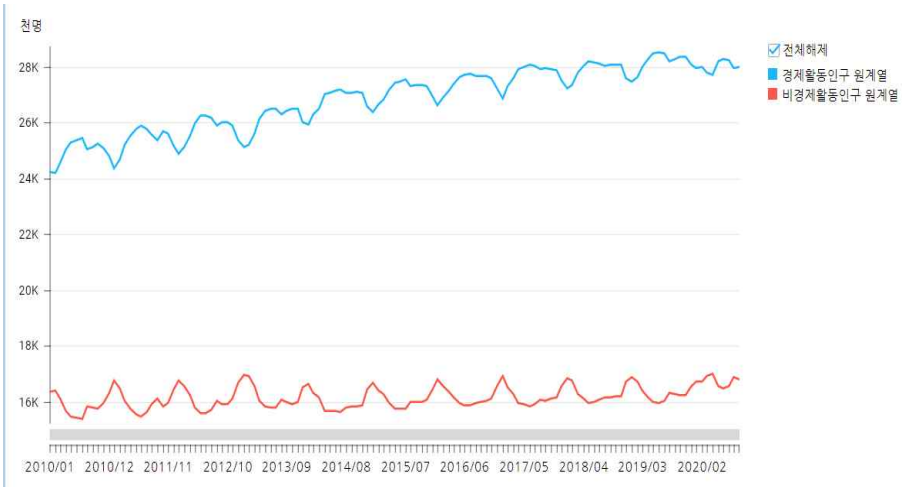
결국 2014년, 취업자의 마음은 대기업보다 중견기업으로 바뀌었고 이는 현실적인 합격 가능성과 자신의 성장 가능성을 바탕으로 목표 기업을 선택하여 실제 기업의 지원자가 증가하는 효과로 이어졌습니다. [출처 인용]
이 효과는 2010년대에서 2014년 비경제활동인구수가 가장 낮은데 주요한 원인이 되었다고 추정됩니다. [출처 인용]

2020년은 '코로나 바이러스'라는 불규칙 성분이 강타한 이후 많은 것이 바뀌었습니다. 특히 2020/6 월, 계절패턴을 유지한다면 낮은 비경제활동인구를 기록할거라 예측되었지만 가장 높은 인구수를 보여준 2019/01 월과 거의 비슷했습니다. 이는 2020년 코로나로 인한 채용시장의 대폭감소 및 사라진 기업의 증가가 주요 원인으로 작용하여 취업에 대한 마음가짐에 영향을 준 것으로 보입니다.

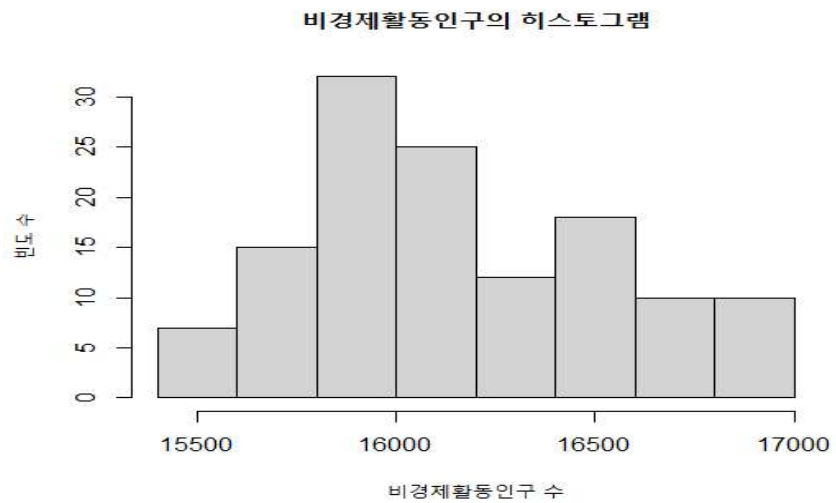
3. 본문

(기본) 이용한 시계열 자료 설명 - 데이터의 주요 특징

시도표

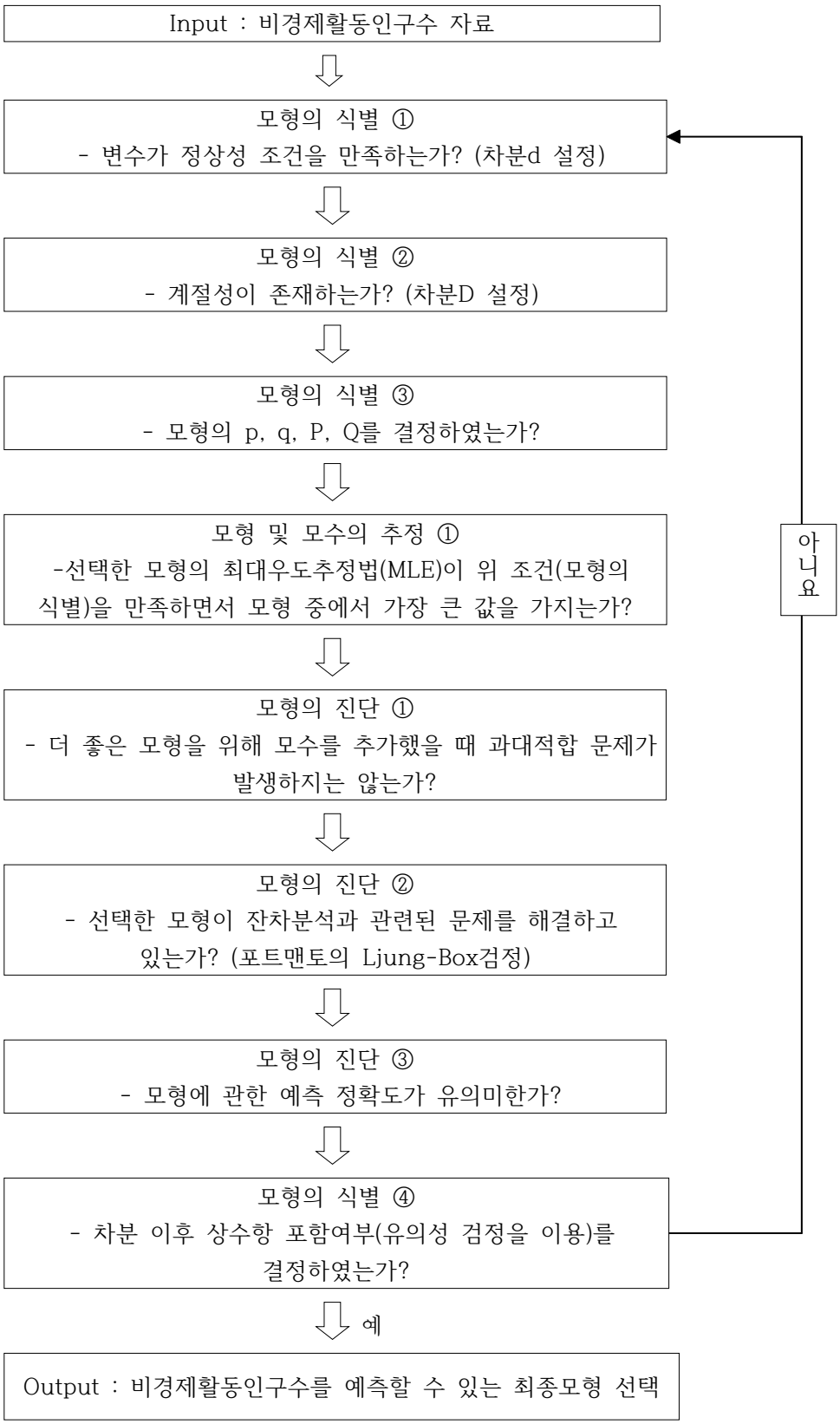


히스토그램



정상성 : 정상성을 만족시키지 못하는 비정상시계열의 형태를 띄고 있습니다.
계절성 : 1년을 기준으로 비슷한 형태의 그래프가 이어지고 있습니다.
특히, 1월에 가장 높은 수치를 보이고 6월에 가장 낮은 수치를 보입니다.
차분 : 계절성 차분이 필요해 보이고 ACF/PACF 그래프를 그려본 뒤 차분 이후에도 정상성을 만족하지 못한다면 일반 차분 여부를 결정하고자 합니다.

3-0. ARIMA 모형 적합의 상세절차



=> 보통 제공되는 간단한 프로세스 (모형식별 -> 모형 및 모수 추정 -> 모형 진단)에서 좀 더 구체적으로 프로세스를 작성해보았습니다.

그럼 두 가지 방법으로 프로그램을 돌릴 수 있었습니다.

첫째로 프로세스를 진행할 때 모형의 식별단계에서 대량의 모형을 만들어서 모형의 추정 방향으로 내려가는 방법을 사용하기도 했습니다.
조건을 만족하는 모형은 내려가고 만족하지 않은 모형은 제거하는 방식으로 진행할 수 있었습니다.

두 번째로 하나의 모형을 만들고 조건을 만족하지 않으면 다시 돌아가 새로운 모형을 만드는 방법이었습니다.
그 결과 화살표 아래로 내려가는 프로세스 즉, 단방향이 아닌 쌍방향 화살표로 진행하는 방법도 좋았던 것 같습니다.

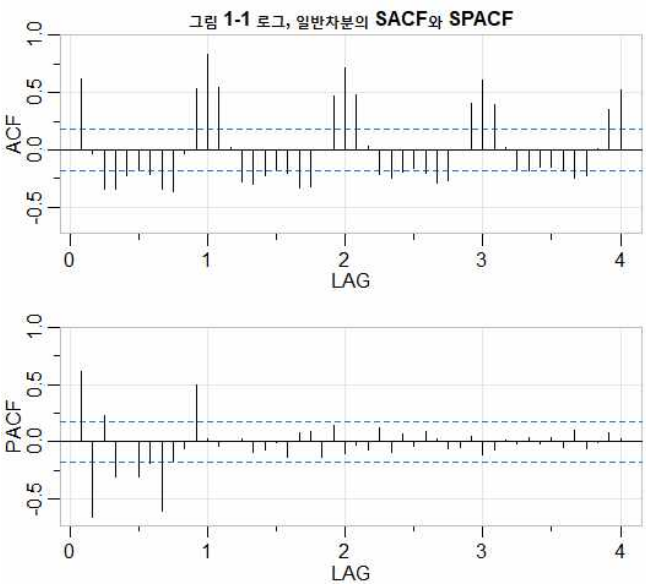
3-1. 모델 식별 및 p, d, q 선택

3-1-1) 변수가 **정상성**을 만족하는지 확인합니다.
3-1-2) 종속 계열의 **계절성**을 식별한다. (필요한 경우 계절성 차분을 진행합니다.)
참고. 시계열의 특징(T, C, S, I)을 확인 후 적절한 **차분**을 실시합니다.
3-1-3) **ARMA모형의 (p,q)**를 결정합니다. 이때 종속 시계열의 자기상관성(ACF) 및 부분 자기상관성(PACF) 함수의 그래프를 사용하여 모델에 어떤 자동회귀(AR) 또는 이동 평균(MA) 구성 요소를 사용해야 하는지 결정한다.
3-1-4) 만약 차분을 진행하였다면 **상수항 포함 여부**를 결정합니다.

=> 대전제 : **모수 간결의 원칙**을 생각한다. (모수 간결의 원칙 : 모수의 개수가 많아지면 최종 예측모형이 복잡해질 뿐만 아니라 추정의 효율성도 떨어지므로 가능한 간단한 모형을 선택) [위키백과]

3-1-1) 그래프를 그려보면 변수인 비경제활동인구수가 정상성을 만족하지 않는 모습을 보입니다, 따라서 비정상시계열인 비경제활동인구 수 데이터를 **로그변환**과 **일반차분**을 통해 정상성을 만족시키고자 합니다.

========== [R코드] ==========
```
>d1lpop<-diff(lpop, lag=1) # d=1 일반 차분
>acf2(d1lpop, main = "그림 1-1 로그, 일반차분의 SACF와 SPACF")
```

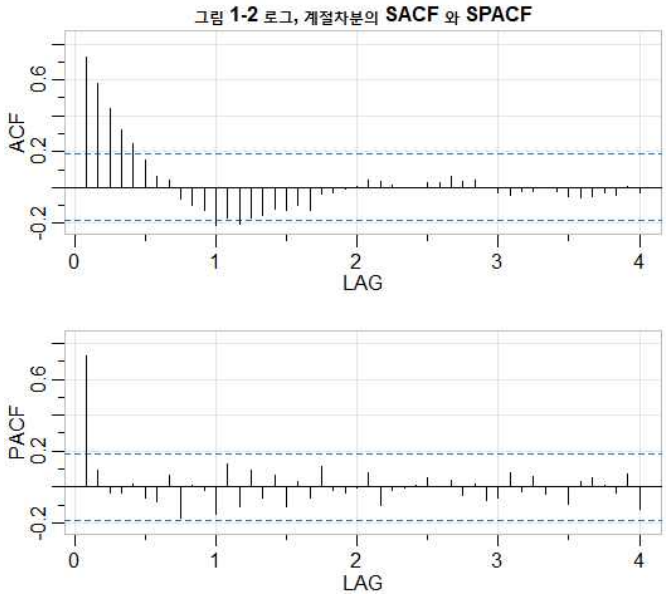


==========

3-1-2) 그래프가 월별 계절성을 띄어 **계절성 차분**을 진행하였습니다.

============================== [R코드] ==============================

```
>d12lpop <- diff(lpop, lag = 12) #계절차분
>acf2(d12lpop, main="그림 1-2 로그, 계절차분의 SACF 와 SPACF")
```



==========

- 차분이 필요한 경우 -

① 추세나 계절성이 존재할 때 (O)

1번 : 추세 X, 계절성 O -> 계절차분만 진행

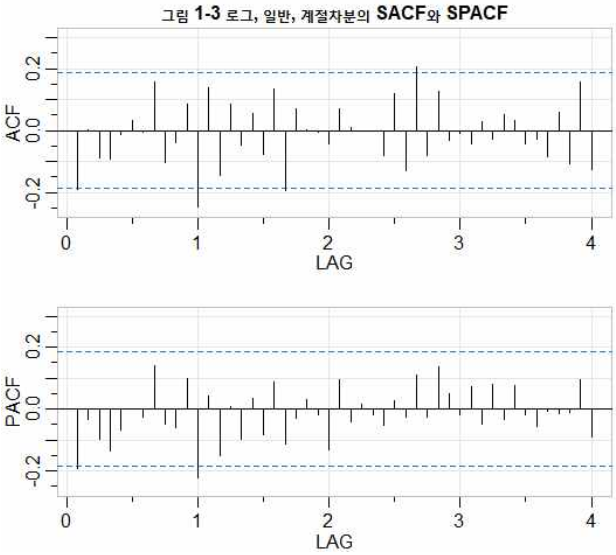
2번 : 추세 O, 계절성 X -> 1차 차분(선형), 로그변환+1차 차분(곡선)

3번 : 추세 O, 계절성 O -> 계절차분 + 추세검토 후 1차 차분 진행
(과대차분을 피하기위해)

-> 위 사항에서 1번과 2번을 모두 시행했지만 정상성 조건을 만족하지 못해 3번인 **계절차분 + 추세검토 후 1차 차분** 진행을 진행하고자 합니다.

============================== [R코드] ==============================

```
>d1_12lpop <- diff(d12lpop) # 일반 1차차분과 계절차분 모두시행
>acf2(d1_12lpop, main = "그림 1-3 로그, 일반, 계절차분의 SACF와 SPACF")
```



==========

=> **정상성** 조건을 만족하는 것을 볼 수 있다.

② 차분된 시계열(계절+일반 차분)의 분산이 이전(일반차분)의 분산보다 작은 경우 (O)
로그변환과 일반차분만 진행한 경우

```
>arima(lpop, order = c(0,1,1))
[1] sigma^2 estimated as 0.0001776
```

로그변환과 일반차분, 계절차분까지 진행한 경우

```
>sarima(lpop, 0,1,1, 1,1,0, 12)
[1] sigma^2 estimated as 5.914e-05
```

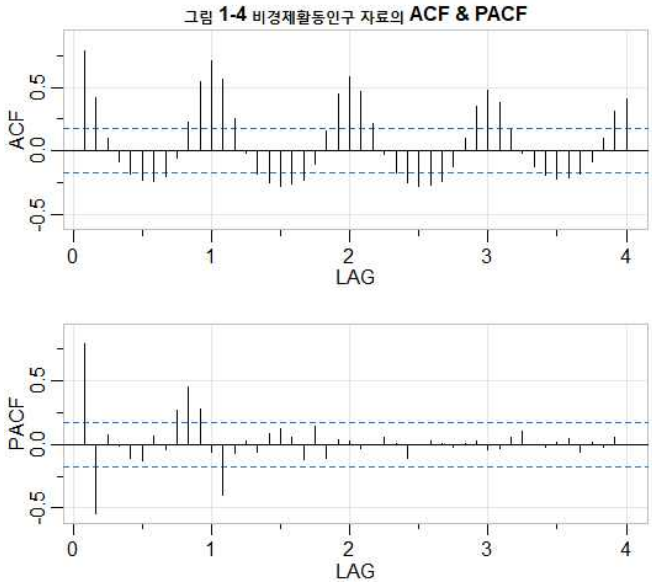
3-1-3) ARMA모형의 (p,q)를 결정하는 방법

① **ACF와 PACF**를 확인하는 방법

=> ACF와 PACF 그래프를 확인해보고자 합니다. 이때, AR모형에서의 그래프가 점진적인 감소가 아닌 지수적 감소가 일어난다면 좋은 모형이라 할 수 있겠습니다. 하지만 비경제활동인구수의 ACF/PACF는 점진적으로 감소하는 모습을 보여줍니다.

============================== [R코드] ============================

>acf2(pop, main ="그림 1-4 비경제활동인구 자료의 ACF & PACF")



==========

앞에서 차분을 진행했기 때문에 arima 모형을 결정짓고 p, q 값을 구하고자 합니다. 그런데 이 경우에는 계절성분을 가지고 있기 때문에 sarima 모형을 적용시키고자 합니다. 즉, sarima(p, 1, q)(P, 1, Q)(12)로 토대를 정했고 모수 간결의 원칙을 적용하여 p, q, P, Q에 1씩 늘어나는 방향으로 모형을 만들려고 합니다.

②AIC, AICc, BIC 값의 비교를 통한 방법

=> AIC, AICc, BIC 값의 비교를 통한 방법은 모형을 만든 것 중에서 값이 작은 모형은 선택하고자 합니다. 많은 모형들이 있지만 일단 sarima(0,1,1)(0,1,1)(12) 모형과 sarima(0,1,1)(1,1,0)(12) 모형을 비교해보고자 한다.

============================== [R코드] ============================

1번 : sarima(0,1,1)(0,1,1)(12) 모형

```
>fit333 = arima(lpop, order = c(0,1,1),
                seasonal = list(order = c(0,1,1), period =12),
                include.mean = F); fit333
```

Call:

arima(x = lpop, order = c(0, 1, 1), seasonal = list(order = c(0, 1, 1), period = 12))

Coefficients:

	ma1	sma1
	-0.1945	-0.5466
s.e.	0.1166	0.1183

sigma^2 estimated as 5.554e-05: log likelihood = 401.56, aic = -797.13

==========

=> $(1-B)(1-B^{12})\ln Z_n$

$= (1-0.1945B)(1-0.5466B^{12})\ln\epsilon_t$

========== [R코드] ==========

2번 : sarima(0,1,1)(1,1,0)(12) 모형

```
>fit111 <- arima(lpop, order = c(0,1,1),
                 seasonal = list(order = c(1,1,0), period =12),
                 , include.mean = F); fit111
```

Call:

arima(x = lpop, order = c(0, 1, 1), seasonal = list(order = c(1, 1, 0), period = 12))

Coefficients:

	ma1	sar1
	-0.2261	-0.3706
s.e.	0.1151	0.1087

sigma^2 estimated as 5.914e-05: log likelihood = 399.16, aic = -792.31

==========

=> $(1-0.37063B^{12})\ln Z_t = (1-0.22614B)\ln\epsilon_t$

둘 중에서 sigma^2와 aic는 작으면 좋고 log likelihood는 크면 좋다. 따라서.sarima(0,1,1)(0,1,1)(12) 모형을 선택하고자 합니다. 이 이상 p나 q를 늘리게 되면 모수 간결의 원칙에 위배되어 일단 arima(0, 1, 1)(0, 1, 1)(12) 모형을 선택하고자 합니다.

하지만 다음 단계인 차분 실행 후 상수항 포함 여부 과정을 거치면서 선택한 모형의 계수의 유의성에 문제가 생깁니다. 즉, 유의성 검정에서 통과하지 못하는 모습을 보여주어 최종적으로 arima(0,1,1)(1,1,0)(12) 모형을 선택합니다.

3-1-4) **차분**을 실행한 경우

상수항을 모형에 포함시킬지 여부는 유의성 검정을 통해 결정한다. 하지만 이 경우에서는 계절차분과 일반차분이 모두 이루어져있기 때문에 상수항은 자동적으로 제거된 형태를 가진다. 이를 확신하기 위해 실제로 R코드를 돌려봤을 때 같은 숫자가 나옴을 볼 수 있다.

```
============================== [R코드] ==============================
- 상수항을 포함하지 않는 경우 -
>fit333 <- arima(lpop, order = c(0,1,1),
+                 seasonal = list(order = c(0,1,1), period =12),
+                 , include.mean = F); fit333
>coeftest(fit333)

z test of coefficients:

     Estimate Std. Error z value  Pr(>|z|)
ma1  -0.19450    0.11663 -1.6677    0.09538 .
sma1 -0.54659    0.11831 -4.6200 3.837e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
============================== [R코드] ==============================

- 상수항을 포함하는 경우 -
> fit333 <- arima(lpop, order = c(0,1,1),
+                 seasonal = list(order = c(0,1,1), period =12)); fit333
 coeftest(fit333)

z test of coefficients:

     Estimate Std. Error z value  Pr(>|z|)
ma1  -0.19450    0.11663 -1.6677    0.09538 .
sma1 -0.54659    0.11831 -4.6200 3.837e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
=================================================================
=> 유의성 검정을 만족하지 않는다.
그래서 sarima(0,1,1)(1,1,0)(12) 모형을 선택하여 다시 진행합니다.
```

============================== [R코드] ==============================

- 상수항을 포함하지 않는 경우 -

```
>fit111 <- arima(lpop, order = c(0,1,1),
                 seasonal = list(order = c(1,1,0), period =12),
                 , include.mean = F); fit111
>coeftest(fit111)
```

z test of coefficients:

```
     Estimate Std. Error z value Pr(>|z|)
ma1  -0.22614    0.11515 -1.9639 0.049539 *
sar1 -0.37063    0.10867 -3.4107 0.000648 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- 상수항을 포함하는 경우 -

```
>fit112 <- arima(lpop, order = c(0,1,1),
                 seasonal = list(order = c(1,1,0), period =12)); fit112
>coeftest(fit112)
```

z test of coefficients:

```
     Estimate Std. Error z value Pr(>|z|)
ma1  -0.22614    0.11515 -1.9639 0.049539 *
sar1 -0.37063    0.10867 -3.4107 0.000648 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

==========

3-2. 모형의 추정

3-2-1) 파라미터 추정:
이전 단계에서 여기까지 오면서 선택한 ARIMA 모델에 적합한 계수를 찾는다. 가장 일반적인 방법은 **최대우도추정(MLE)** 또는 조건/비조건부 최소제곱추정법이 있다.

=>
3-2-1) 파라미터 추정 진행방향:
최대우도추정법(MLE)은 최대한 큰 값을 가지는 모형이 좋다고 판단합니다. 따라서 최대우도추정으로 선택된 모형의 계수를 찾는다.

============================== [R코드] ==============================

```
>sarima(lpop, 0,1,1, 1,1,0, 12) #fit111
sigma^2 estimated as 5.914e-05:  log likelihood = 399.16,  aic = -792.31

Coefficients:
          ma1     sar1
      -0.2261  -0.3706
s.e.   0.1151   0.1087
```

==========
=> 즉, 모형의 계수는 ma1은 -0.2261이고 sar1은 -0.3706 입니다.

3-3. 모형의 진단

3-3-1) 과대 적합 : 잠정모형에 모수를 추가하여 더 많은 개수의 모수를 포함하는 모형을 적합시키는 경우 [위키백과]

3-3-2). 잔차검정 및 통계모형, 조건, 포트맨토검정(Ljung-Box test):

3-3-3) 모형에 관한 예측 정확도 평가

=>
3-3. 모형의 진단
3-3-1) 과대적합

잠정 모형 : arima(0,1,1)(1,1,0)(12)

```
============================== [R코드] ========================
> fit111 <- arima(lpop, order = c(0,1,1),
+                 seasonal = list(order = c(1,1,0), period =12)); fit111
> fit111$aic # 더 작은 것이 좋은 모형
[1] -792.3123
> fit111$sigma2
[1] 5.913911e-05
=================================================================
```

새로운 모형 : arima(0,1,2)(1,1,0)(12) 모형 적합

```
============================== [R코드] ========================
> fit111 <- arima(lpop, order = c(0,1,2),
+                 seasonal = list(order = c(1,1,1), period =12)); fit333
>fit333$aic # 더 작은 것이 좋은 모형
[1] -797.8568
> fit333$sigma2
[1] 5.40073e-05
=================================================================
```

두 모형 중에서 aic와 sigma2를 비교하여 모형을 고른다면, 더 작은 값을 가지는 새로운 모형 arima(0,1,2)(1,1,0)(12)을 적합하는 것이 좋지만 과대적합 문제를 해결해야 한다. 하지만 과대적합에 관한 아래의 조건을 만족하지 못하므로 새로운 모형이 아닌 잠정모형을 선택하도록 한다.

- 과대적합이 아니기에 새로운 모형을 쓸 수 있는 조건 -

① 추가모수가 유의하다. (X)

=> ma2가 유의수준0.05에서 유의하지 않는 모습을 보여준다

============================== [R코드] ============================

```
> coeftest(fit1111)
```

z test of coefficients:

```
       Estimate Std. Error z value  Pr(>|z|)
ma1  -0.244912   0.094124 -2.6020 0.0092676 **
ma2  -0.154807   0.095994 -1.6127 0.1068170
sar1 -0.402899   0.106895 -3.7691 0.0001638 ***
```

==========

② 잠정모형의 모수추정값이 과대적합 후 모수추정값과 큰 차이를 보인다. (X)

=> 기존 ma1 = -0.2261, sar1= -0.3706

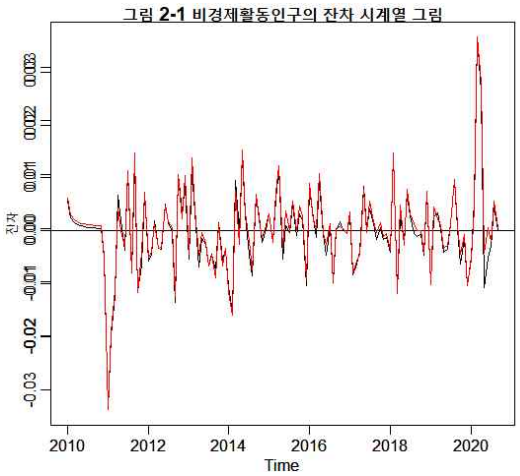
변화 ma1 = -0.244912, ma2 = -0.154807, sar1 = -0.402899

③ 과대 적합모형의 잔차분산이 잠정모형보다 작아진다. (X)

=> 그래프로 비교하고자 한다. 검은 선은 잠정모형의 잔차, 빨간 선은 새로운 모형의 잔차를 나타낸다. 결과를 보면 대부분의 선이 겹쳐져 있는 모습을 보여주기에 잔차분산이 작아졌다고 보기는 힘들다.

============================== [R코드] ==============================

```
>ts.plot(resid(fit111), ylab = "잔차", main = "그림 2-1 비경제활동인구의 잔차 시계열 그림"); abline(h=0)
>par(new=T)
>ts.plot(resid(fit11), ylab= "", col='red')
```



==========

이때 **모수과잉**을 방지하기 위해 AR과 MA를 동시에 추가하지 않는다

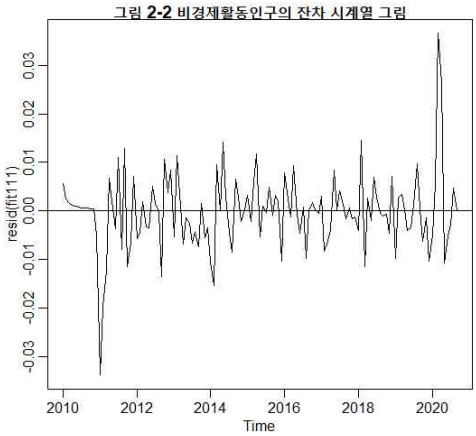
=> 두 개의 모수가 동시에 추가될 경우 공통요인이나 거의 공통인 요인이 모형의 AR과 MA에 존재할 수 있어 공통항을 서로 상쇄시켜줘야지만 모수과잉 현상이 일어나지 않음

3-3-2). 잔차검정 및 통계모형, 조건, 포트맨토검정(Ljung-Box test)의 진행

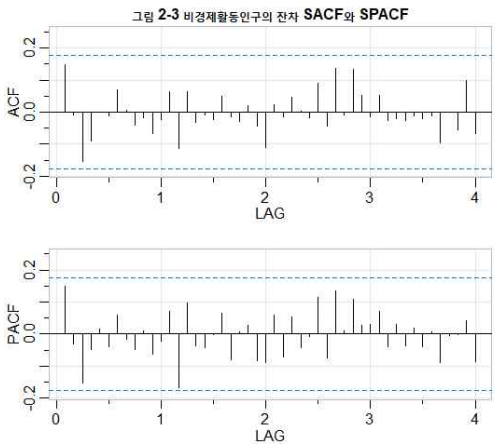
잔차의 정규성

========== [R코드] ==========

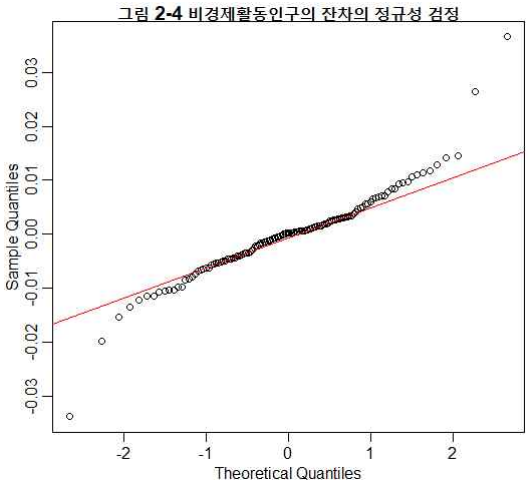
>ts.plot(resid(fit111), main = "그림 2-2 비경제활동인구의 잔차 시계열 그림"); abline(h=0)



>acf2(resid(fit111), maxlag=24, main="그림 2-3 비경제활동인구의 잔차 SACF와 SPACF")



>qqnorm(resid(fit111), main = "그림 2-4 비경제활동인구의 잔차의 정규성 검정")
>qqline(resid(fit111), col="red")



==========

=> 잔차들이 정규분포에 근사하는지를 확인합니다. 즉, 직선에 점들이 가까울수록 정규성이 있다고 판단합니다. 따라서 위 그림을 보면 정규성을 잘 만족하고 있다. 직선을 벗어나는 점들은 잔차시계열 그림에서의 **이상치**들 입니다.

- 잔차의 포트맨토 검정 -

========== [R코드] ==========

```
> Box.test(resid(fit111), lag=6, type="L", fitdf = 2) #ljung-box test
        Box-Ljung test

data:  resid(fit111)
X-squared = 7.2585, df = 4, p-value = 0.1228

> Box.test(resid(fit111), lag=12, type="L", fitdf = 2) #ljung-box test
        Box-Ljung test

data:  resid(fit111)
X-squared = 8.9868, df = 10, p-value = 0.5334

> Box.test(resid(fit111), lag=24, type="L", fitdf = 2) #ljung-box test
        Box-Ljung test

data:  resid(fit111)
X-squared = 15.211, df = 22, p-value = 0.853
```

==========

Ljung-Box 검정의 가설 설정

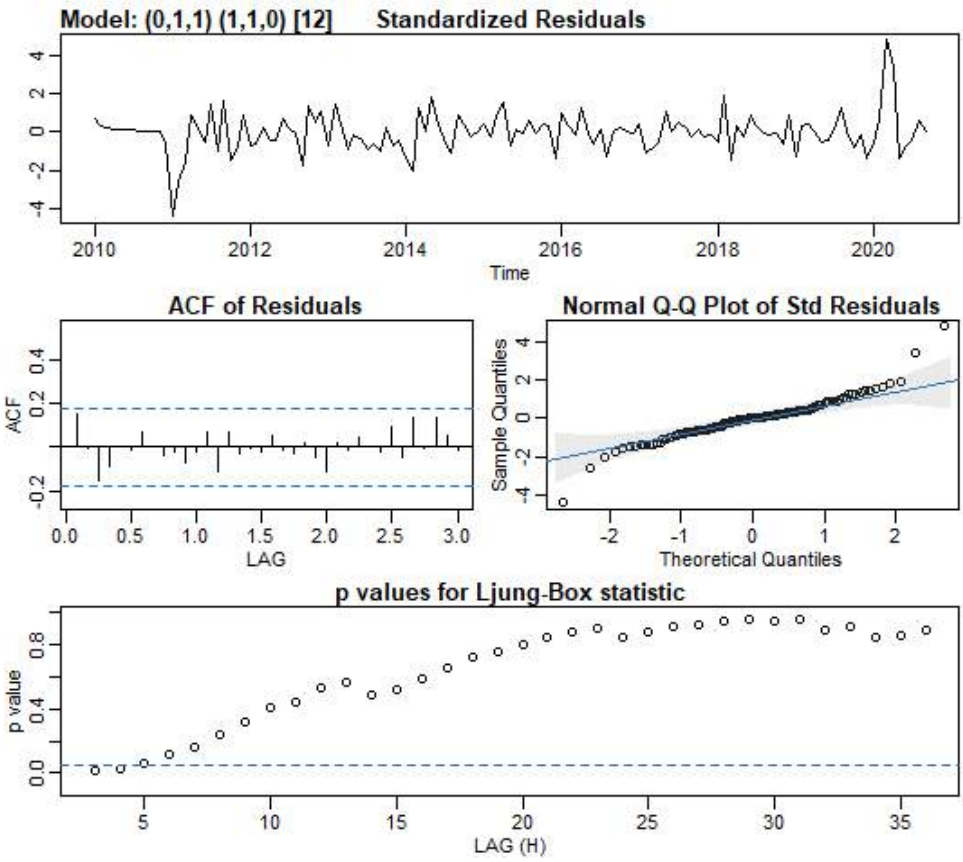
H_0 : 데이터가 독립적으로 분산된다. (즉, 시차 1에서 k까지의 모집단의 **상관관계는 0**이므로 데이터에서 관찰 된 상관관계는 표본의 무작위성에서 기인합니다)

H_1 : 데이터가 독립적이지 못합니다. (자기상관관계를 갖는다, 상관계수가 0이 아니다.)

위에서 볼 수 있듯이 p-value 값이 유의수준 0.05를 넘지 못하므로 귀무가설을 기각하지 못하는 모습을 보여줍니다. 하지만 이는 예측 기법이 개선될 수 없다고 볼 수 없습니다. 즉, 같은 데이터에 대해서 이러한 특징을 모두 만족하는 몇 가지 예측 기법이 있을 수 있습니다.

카이제곱분포를 따르는 큰 Q값은 자기상관계수의 값이 백색잡음 시계열에서 온 게 아닙니다. 위의 결과에서 잔차(residual)가 백색잡음 시계열과 다르진 않다고 결론 내릴 수 있습니다.

마지막으로 sarima(0,1,1)(1,1,0)(12)를 호출해 나오는 그래프를 해석하고자 합니다.



Standardized Residuals 그래프는 arima 모형을 적합시키고 남은 잔차의 성분을 보여줍니다. 이때 남은 잔차가 백색잡음에 가까울수록 좋은 모형이라고 판단합니다. 위 그래프를 보면 2011년과 2020년에서 비정상적으로 큰 부분이 존재하지만 백색잡음에 가깝다고 볼 수 있습니다.

두 번째로 Normal Q-Q plot of Std Residuals에서는 잔차들이 정규분포에 근사하는지를 확인합니다. 직선에 점들이 가까울수록 정규성이 있다고 판단합니다. 위 그래프는 직선과 예측범위 안에 대부분의 점이 존재하고 몇 개의 점이 바깥쪽에 있기에 잔차가 정규성을 만족한다고 해석합니다.

맨 밑의 p value for Ljung-Box statistic 그래프는 점들이 파란선 위에 있을수록 좋은 모형이라고 판단합니다. 위 그래프는 몇몇 점들이 p-value를 초과하지 않는 모습을 보여주기에 좋은 모형이라고 생각했습니다.

3-3-3) 모형에 관한 예측 정확도 평가
=> 각종 잔차에 관련된 예측 정확도를 보면 충분히 적은 값들을 가지고 있다. 즉, 선택한 모형이 크게 부적절하다고 볼 수 없겠다.

============================== [R코드] ==============================

```
> summary(fit111)

Training set error measures:
                      ME       RMSE          MAE          MPE       MAPE
Training set -0.000156364 0.0079091 0.005377474 -0.00162352 0.05544435
                   MASE       ACF1
Training set 0.4646172 0.1488422
```

==========

3-4. 최종모형에 대한 해석(계수의 유의성)

```
============================== [R코드] ==============================
> coeftest(fit111)
z test of coefficients:

     Estimate  Std. Error z value Pr(>|z|)
ma1  -0.22614     0.11515 -1.9639 0.049539 *
sar1 -0.37063     0.10867 -3.4107 0.000648 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
=====================================================================
```

최종모형 : $(1-0.37063B^{12})\ln Z_t = (1-0.22614\,B)\ln\epsilon_t$

최종모형의 선택이유 : 차분이 1인 상태로 고정시키고 p,q,P,Q의 조합을 모두 작성해보았다.

```
1번 : fit111 <- arima(lpop, order = c(0,1,1),
                seasonal = list(order = c(1,1,0), period =12)); fit111
2번 : fit222 = arima(lpop, order = c(1,1,0),
               seasonal = list(order = c(1,1,0), period =12)); fit222
3번 : fit333 = arima(lpop, order = c(0,1,1),
               seasonal = list(order = c(0,1,1), period =12)); fit333
4번 : fit444 = arima(lpop, order = c(1,1,0),
               seasonal = list(order = c(0,1,1), period =12)); fit444
```

1번	2번	3번	4번
sigma^2 estimated as 5.914e-05:	sigma^2 estimated as 5.965e-05:	sigma^2 estimated as 5.554e-05:	sigma^2 estimated as 5.584e-05:
log likelihood = 399.16,	log likelihood = 398.7,	log likelihood = 401.56,	log likelihood = 401.21,
aic = -792.31	aic = -791.41	aic = -797.13	aic = -796.42

다음으로 AIC, AICc, BIC에 관한 내용이다.
1번 : sarima(lpop, 0,1,1, 1,1,0, 12) #fit111
2번 : sarima(lpop, 1,1,0, 1,1,0, 12) #fit222
3번 : sarima(lpop, 0,1,1, 0,1,1, 12) #fit333
4번 : sarima(lpop, 1,1,0, 0,1,1, 12) #fit444

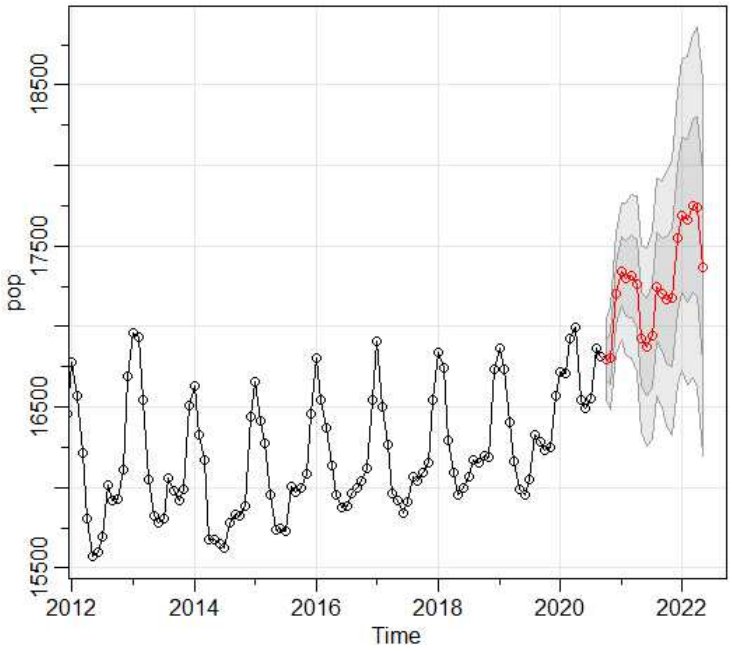
1번	2번	3번	4번
$AIC [1] -6.23868	$AIC [1] -6.231546	$AIC [1] -6.276581	$AIC [1] -6.271031
$AICc [1] -6.237918	$AICc [1] -6.230784	$AICc [1] -6.275819	$AICc [1] -6.270269
$BIC [1] -6.173634	$BIC [1] -6.1665	$BIC [1] -6.211535	$BIC [1] -6.205985

이를 보았을 때
3번 모형 > 4번 모형 > 1번 모형 > 2번 모형 순으로 좋은 모형입니다.
하지만 3번 모형 $(1-B)(1-B^{12})\ln Z_n$에서 모수 유의성 검정을 시행했을 때, 유의수준 0.05를 만족하지 않는 것을 보여주었습니다. 4번 모형 역시 마찬가지였습니다. 따라서 1번 모형인
$(1-0.37063B^{12})\ln Z_t = (1-0.22614\,B)\ln\epsilon_t$을 최종모형으로 선택했습니다.

최종 모형의 해석 : 먼저 최종모형에서 주기가 12인 계절 1차 차분 시계열이 AR(1)을 따르고 비계절형 1차 차분 시계열은 MA(1)을 따릅니다. 그리고 계수의 유의성에서 보자면, 유의수준 0.05에서 ma1과 sar1은 모두 유의한 모습을 보입니다.
p, q를 결정한 것 이외에도 차분인 d의 수를 늘리는 것도 생각해보았습니다. 하지만 [그림 1-3]에서 보았듯이 이미 정상성을 만족하고 있는데 차분을 2로 늘릴 이유가 없습니다. 또한 과대차분이 일어나 ACF/PACF를 복잡하게 만들거나 분산이 커지는 문제도 발생할 우려가 있어서 차분을 1로 고정하는 모형을 선택하였습니다.

3-5. 미래 20개 시점에 대한 예측과 평가

여러 예측방법 중에서 최소제곱오차(MMSE)를 이용한 예측을 하려고 합니다.
비경제활동인구 데이터에서 미래 20개 시점은 2020년 10월부터 2022년 5월까지를 나타냅니다.



예측값과 예측오차

============================== [R코드] ==============================

```
> sarima.for(pop, 20, 0,1,1, 1,1,0, 12 )
```

$pred 예측값

```
          Jan       Feb       Mar       Apr       May       Jun       Jul
2020
2021 17344.97 17294.50 17312.72 17265.78 16922.80 16872.26 16945.81
2022 17692.14 17658.21 17748.01 17742.87 17363.52
          Aug       Sep       Oct       Nov       Dec
2020                    16794.50 16802.77 17203.79
2021 17246.81 17200.53 17165.48 17176.93 17548.84
2022
```

$se 예측오차

```
          Jan       Feb       Mar       Apr       May       Jun       Jul
```

2020
2021 212.5856 234.4116 254.3717 272.8756 290.2021 306.5508 322.0707
2022 483.3917 509.6942 534.7044 558.5959 581.5067
          Aug   Sep   Oct   Nov   Dec
2020   126.1911 160.2509 188.2458
2021 336.8764 351.0582 394.0884 425.9416 455.5732
2022

==========

예측평가

2020년 10월부터 2022년 5월까지의 예측을 보면, 코로나로 인한 비경제활동인구수가 증가하는 모습으로 예측되었습니다.

한 달 뒤인 2020년 10월 데이터가 업데이트되어 비교해보자면 예측값은 16794.5명, 실제 데이터는 16736명으로 빨간 예측선이 정확하게 맞추진 못했지만 꽤 근접하게 예측했음을 볼 수 있었습니다. 더하여 파란 범위까지 본다면 충분히 범위 안에 들어가는 것을 볼 수 있었습니다.

하지만 파란 예측오차범위는 시간이 지남에 따라서 그 예측범위가 넓어지는 모습을 알 수 있습니다. 그래서 일반적으로 빨간 예측선을 중심으로 보는 것이 아닌 파란 예측범위를 중심으로 보면 미래에 대한 예측이 설득력이 있느냐는 문제가 제기될 수도 있겠습니다.

또한 가장 중점적으로 봐야 할 것은 시계열 분석에서 빨간 예측선은 물론이고 파란색 예측오차의 범위까지 포함하더라도 다시 계절 패턴을 유지할 거로 예측한다는 사실입니다. 이는 현재 가지고 있는 데이터를 바탕으로 한다면 정당하면서 최적의 예측이라 생각합니다. 하지만 간과한 부분이 있어 잘못된 예측이 될 수도 있을 거란 생각이 들었습니다.

먼저 사람은 '코로나 바이러스'라는 구체적인 원인을 알고 있어서 불규칙적인 성분임을 알 수 있었습니다. 하지만 컴퓨터는 데이터만을 가지고 정확한 이유는 모르지만 불규칙변동으로 판단했습니다. 여기서 문제점은 일차적으로 '코로나 바이러스'를 불규칙변동으로 예측했지만 2020/12월 기준 코로나 바이러스가 다시 크게 유행하면서 단순한 불규칙성분으로 볼 수 있느냐는 것입니다. 여기서는 2022년 5월까지 예측했지만 더 긴 시점을 예측한다면 강한 추세 패턴이 생길 가능성이 있다고 생각합니다. 따라서 불규칙 변동으로 인해 계절패턴을 다시 유지할 것이라는 예측은 최고의 정답이 아닐 수도 있다고 생각해야 합니다.

범위를 넓혀 다시 말하자면 시계열 분석의 예측은 항상 정답이 아닌 미래결과의 보기 중 하나라는 것을 명심해야 할 것입니다.

4. 별첨

4-1. 비경제활동인구의 내용 분석의 참고자료

<참고자료>
[2013 취업시장은? 고졸채용 증가·채용규모 양극화]
[https://www.ajunews.com/view/20131203125555267]
["사람인, 2014 취업시장 결산!"]
[http://www.saramin.co.kr/zf_user/help/live/view?idx=24884&listType=news]

4-2. R-Code

```
library(astsa)
library(lmtest)
library(forecast)
library(lubridate) #ymd function
library(portes) #ljungbox
library(fUnitRoots) #adfTest function 사전에 패키지
library(ggplot2)

z<- scan("C:/Users/Playdata/Documents/r1.txt") # 데이터 호출
pop <- ts(z, start = c(2010, 1), frequency = 12) # 시계열 데이터로 변환
pop

lpop<-log(pop) # 로그변환
d1lpop<-diff(lpop, lag=1) # d=1 일반 차분
acf2(d1lpop, main = "그림 1-1 로그, 일반차분의 SACF와 SPACF")

d12lpop <- diff(lpop, lag = 12) #계절차분
acf2(d12lpop, main="그림 1-2 로그, 계절차분의 SACF 와 SPACF")

d1_12lpop <- diff(d12lpop) # 일반 1차차분과 계절차분 모두시행
acf2(d1_12lpop, main = "그림 1-3 로그, 일반, 계절차분의 SACF와 SPACF")

arima(lpop, order = c(0,1,1)) # 비계절성 MA(1)과 1차 차분
```

```
sarima(lpop, 0,1,1, 1,1,0, 12) #fit111와 동일

acf2(pop, main ="그림 1-4 비경제활동인구 자료의 ACF & PACF")

#계절성은 AR(1)모형, 비계절성은 MA(1)모형을 따르고 각각 1차 차분 실행
fit111 <- arima(lpop, order = c(0,1,1),
                seasonal = list(order = c(1,1,0), period =12),
                include.mean = F); fit111

#계절성은 AR(1)모형, 비계절성은 AR(1)모형을 따르고 각각 1차 차분 실행
fit222 = arima(lpop, order = c(1,1,0),
               seasonal = list(order = c(1,1,0), period =12),
               include.mean = F); fit222

#계절성은 MA(1)모형, 비계절성은 MA(1)모형을 따르고 각각 1차 차분 실행
fit333 = arima(lpop, order = c(0,1,1),
               seasonal = list(order = c(0,1,1), period =12),
               include.mean = F); fit333

#계절성은 MA(1)모형, 비계절성은 AR(1)모형을 따르고 각각 1차 차분 실행
fit444 = arima(lpop, order = c(1,1,0),
               seasonal = list(order = c(0,1,1), period =12),
               include.mean = F); fit444

sarima(lpop, 0,1,1, 1,1,0, 12) #fit111와 동일
sarima(lpop, 1,1,0, 1,1,0, 12) #fit222와 동일
sarima(lpop, 0,1,1, 0,1,1, 12) #fit333와 동일
sarima(lpop, 1,1,0, 0,1,1, 12) #fit444와 동일

coeftest(fit111) # 각 계수의 유의성 파악
coeftest(fit222)
coeftest(fit333)
coeftest(fit444)

summary(fit111) # 계수값, 분산, 우도, aic, 오차값 파악
summary(fit222)
summary(fit333)
```

```
summary(fit444)

fit111$aic # aic 값을 파악
fit222$aic
fit333$aic
fit444$aic

fit111$sigma2 # 분산값을 파악
fit222$sigma2
fit333$sigma2
fit444$sigma2

fit1111 = arima(lpop, order = c(1,2,0),
                seasonal = list(order = c(0,2,1), period =12),
                include.mean = F); fit1111 #과대차분을 시도

fit11 <- arima(lpop, order = c(0,1,2),
               seasonal = list(order = c(1,1,0), period =12),
                include.mean = F); fit11 #과대적합을 시도


Box.test(resid(fit111), lag=6, type="L", fitdf = 2) #ljung-box test
Box.test(resid(fit111), lag=12, type="L", fitdf = 2) #ljung-box test
Box.test(resid(fit111), lag=24, type="L", fitdf = 2) #ljung-box test

ts.plot(resid(fit111), ylab = "잔차", main = "그림 2-1 비경제활동인구의 잔차 시
계열 그림"); abline(h=0)
par(new=T) #잔차의 분산을 비교하고자 두 그래프를 겹침
ts.plot(resid(fit11), ylab= "", col='red')
ts.plot(resid(fit111), main = "그림 2-2 비경제활동인구의 잔차 시계열 그림");
abline(h=0)
acf2(resid(fit111), main="그림 2-3 비경제활동인구의 잔차 SACF와 SPACF")

qqnorm(resid(fit111), main = "그림 2-4 비경제활동인구의 잔차의 정규성 검정")
qqline(resid(fit111), col="red")

sarima.for(pop, 20, 0,1,1, 1,1,0, 12) # 미래 20시차 예측
```

4-3. 비경제활동인구 데이터의 출처

==========

자료명 : 비경제활동인구
자료출처 : 한국은행
자료기한 : 2010년 1월 ~ 2020년 9월 + 2020년 10월(업데이트)
(데이터 자체는 1982년부터 있습니다.)

자료주소 [https://ecos.bok.or.kr/jsp/vis/keystat/#/detail]