

[예측방법론]

제목 : 비경제활동인구를 이용한 분석과 예측을 통한 인사이트 제시
응용통계학과
32142221-서근태

1. 요약(summary) : 분석 및 주요결과를 요약하여 기술
2. 본문 : 분석방법 및 과정, 결과를 구체적 서술

2.1. 개요

2.2 추세분석

2.2.1 추세, 계절성분을 동시에 가진 경우 : 선형계절추세모형

2.3 평활법

2.3.1 윈터스 가법계절 지수평활법

2.4 비교

(별첨) R-code(분석에 사용된 R-script를 보고서에 복사)

1. 요약

1.1. 자료분석의 목적

코로나로 인한 사람들의 취업에 대한 마음가짐(더 공부한다/ 휴식) 등이 반영된 비경제활동인구를 이용하여 저와 같이 한창 취업에 예민한 시기인 사람들에게 한국은행의 비경제활동인구 데이터를 이용하여 분석과 예측을 통해 객관적인 인사이트를 제공해보고자 합니다.

1.2 주요결과

비경제활동인구는 경제활동인구와 거의 반대의 값을 얻기 때문에 참고하려고 합니다. 비경제활동인구의 특징은 대규모 채용이 일어나는 3월과 9월에 가장 적은 인구수를 보여준다. 보통 전반기 채용(3월), 하반기 채용(9월)부터 시작이므로 그 기간 큰 변화없이 어느정도 평할한 모습을 보여준다. 3월부터 채용이 시작되어 확정이 되는 6월에 가장 많은 인구수를 기록했습니다. 2010년부터 보면 매년 6월 가장 낮은 인구수를 기록했습니다.

이와 반대로 12월 ~2월 정도에 가장 많은 인구 수를 보여준다. 이는 보통 매년 말에 행해지는 정년은퇴, 일년을 마무리하는데 필요한 일용직 노동자의 해고와 관련되어 있고 채용시장이 얼어붙은 시기인지라 그만큼 인구 수가 급상승 하는 그래프를 보여줍니다. 2010년부터 보면 매년 1월 가장 많은 비경제활동인구수를 기록했습니다.

그래프에서 큰 변화가 있는 2개의 연도가 있는데 먼저 2013년에는 인구수가 급증하는데 이는 글로벌금융위기가 터지고 중국회사들과의 가격경쟁에서 지면서 많은 기업이 없어진 것이 원인이 된 것 같습니다.

2020년은 코로나 바이러스 라는 불규칙 성분이 강타한 이후 가장 낮은 비경제활동인구를 기록했던 2020/06 월에 가장 높은 2019/01 월과 거의 비슷한 인구수를 보여줍니다. 이는 2020년 코로나로 인한 채용시장의 대폭감소가 주요 원인으로 보여집니다.

짧은 기간 코로나 바이러스가 강타한 곳을 본다면 선형 그래프 적용이 적절해 보이지만 보통 16K에서 크게 변하지 않는 모습을 보여줍니다.

앞으로 예측은 코로나 바이러스와 함께 한다면 인구가 줄어들지 않겠지만 규칙한 그래프가 유지되겠으나 백신 제조 이후엔 다시 원 그래프를 유지할 것으로 기대됩니다.

2.2 추세분석

2.2.1 선형계절추세모형

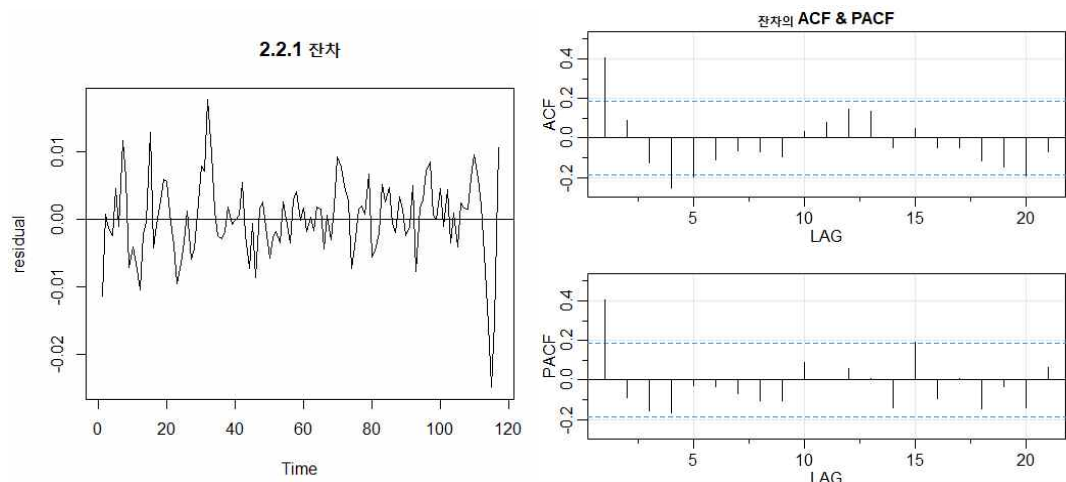
예측모형을 결정한 과정 : 추세성분만 가진 추세모형은 비경제활동인구 그래프에 추세선 하나만 그리게 되어 모형의 예측력이 상당히 낮았고(0.1696), 계절성분만 가진 분석은 추세를 가진 데이터에 적합성이 떨어진다고 보았습니다. 결론적으로 추세, 계절성분을 동시에 가지는 선형계절추세모형을 사용하고자 합니다.

분석방법 및 과정, 결과 : 먼저, 분산의 크기를 줄이고자 기존의 데이터값에 로그를 씌워 데이터를 사용합니다. 이후 로그를 씌운 인구 데이터 $\ln pop$ 를 종속변수로 하고 $0+trend+y$ 를 설명변수로 하는 선형회귀모형 reg 를 제작합니다. y 는 1~12가 반복되는 월별 factor이고 $trend$ 는 추세를 보여주고 형태는 행렬형태로 행은 연도, 열은 월의 형태입니다.

더빈-왓슨 테스트를 보면 p -value가 상당히 낮은 값으로 귀무가설을 기각하지 못하고 자기 상관성이 있다는 것을 알 수 있다. 그리고 $summary$ 를 이용하면 모형의 적합도와 각 변수의 유의성을 알아볼 수 있는데 모형은 1로 잘 적합하다는 것을 알 수 있고 $trend$ 변수와 지시변수가 모두 유의한 모습을 보인다. 이로써 모형식을 얻게 되는데 이는 아래와 같은 값을 가진다.

모형식 : $\ln pop = 0.003t + 9.708INDt1 + \dots + 9.6937INDt12$

잔차와 ACF / PACF 그래프를 보게 되면 0을 중심으로 산재 되어있지 않고 어떠한 형태를 이루면서 자기 상관관계가 있다는 것을 보여준다. 잔차가 독립적이지 못하고 시간이 지남에 따라 패턴이 존재한다. 오차의 등분산의 조건을 벗어났다. 신뢰구간은 파란범위를 나타내는데 넘어가면 자기 상관이 있다는 것을 보여준다. 신뢰구간 안에 있으면 0으로 해석한다



2.3 평활법

2.3.1 윈터스 가법모형

예측모형을 결정한 과정 : 다른 평활법들을 사용해서 예측값을 비교해본 결과 윈터스 가법모형과 승법모형이 그래프로 봤을 때 가장 적합한 모형임을 알게 되었습니다. 그래서 둘을 비교하기 위해 AIC/ AICc/ BIC를 비교하였습니다. AIC는 식 안에 모형의 적합도(likelihood)와 관련되어 크기가 작을수록 올바른 모형에 가깝다고 할 수 있어 더 값이 작은 윈터스 가법모형을 이용합니다.

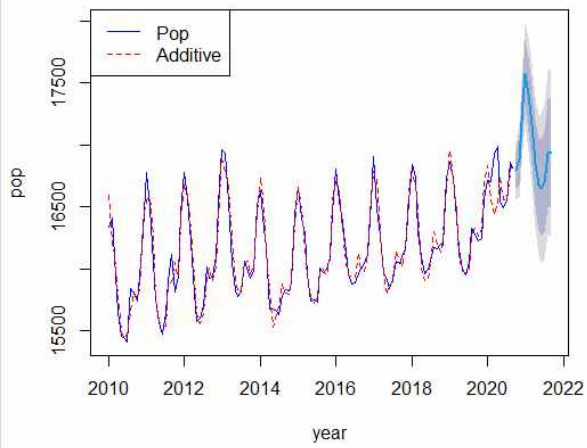
분석방법 및 과정, 결과 : 데이터를 2010년 1월부터 시작하여 12개월로 나뉘고 가법모형을 적용하여 모델의 적합성을 알아봤습니다. 또한 이를 x축 기간 y축을 인구수로 하여 가법모형을 그려보았고 예측오차와 ACF 그래프를 만들었습니다.

모델의 적합성을 보게 된다면, AIC/ AICc/ BIC가 각각 1867.950 / 1882.463 / 1925.566을 얻게 되었습니다. 이는 승법모형인 1900.944/ 1906.458 / 1949.561과 비교하여 더 적은 값을 알 수 있었습니다.

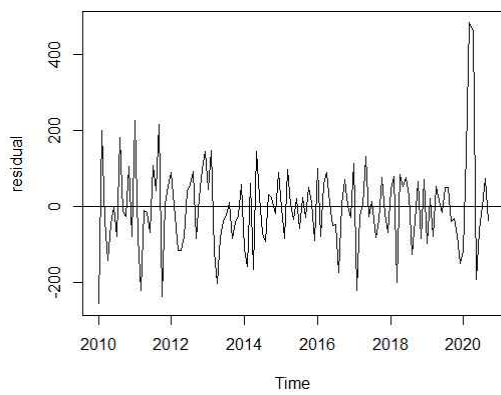
먼저 가법모형 그래프를 보게 된다면, 파란 직선과 빨간 점선이 거의 비슷하게 위치해 있는 것을 볼 수 있습니다. 하지만 코로나가 터진 2020년 그래프는 예상과 달리 실제에서 더 큰 인구수의 증가와 적합되지 않는 모습을 보여줍니다. 또한 이후의 예측이 기존의 그래프를 따르지 않는 것을 보아 큰 폭으로 상승될 것으로 예상되는 그래프를 보여줍니다.

두 번째로 가법모형의 예측오차와 ACF 그래프를 보게 되면 자기 상관이 시점에 따라 남아있는 것으로 보인다. 즉, 체계적인 상관관계를 보입니다. 또한 t-test에서 p-value 값을 보게 되면 0.7918로 귀무가설을 기각하지 못하는 모습을 보여주는데 이는 오차가 랜덤하다.

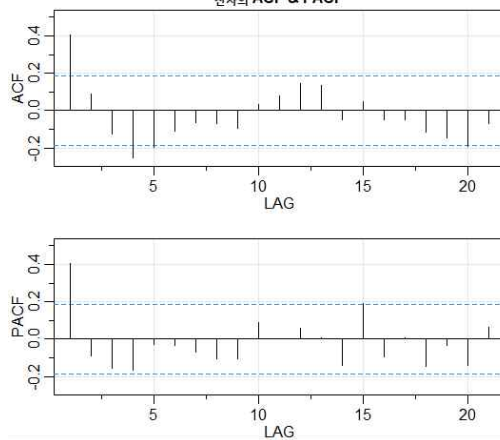
2.3.1 가법모형



2.3.1 가법모형의 예측오차



잔차의 ACF & PACF



2.4 비교

두 모형 다 예측모형과 ACF 그래프를 보면 오차의 등분산의 조건을 벗어난 모습을 보여줬습니다. 하지만 선형계절추세모형은 적합성이 1이 나왔고 윈터스 가법모형의 경우 2020년 전까지 거의 완벽한 예측을 보였습니다. 선형계절추세모형은 모형의 적합성을 볼 수 있어 좋았지만 2020년도 이후의 예측그래프를 볼 수 없어 아쉬웠고 반대로 가법모형은 예측을 볼 수 있어 이후의 활동도 고려해볼 수 있었지만 정확한 값을 가지는 적합성을 볼 수 없어 아쉬웠습니다.

별첨

```
library(astsa) # acf2
library(lmtest) # dwtest
library(forecast)
```

2.2.1 선형계절추세모형

```
z<-read.table("r3.txt", header = F)
pop <- ts(z, frequency = 12, start = c(2010, 1))
lnpop<-log(pop)
trend<- time(lnpop)-2010
trend
y = factor(cycle(lnpop)) #factor를 이용한 월별 factor 생성 / 1~12가 반복
y

reg <- lm(lnpop ~ 0+trend+y) #선형계절추세모형
dwtest(reg)
summary(reg)

ts.plot(pop, fitted(reg), xlab="day", ylab= "sold pop", lty = 1:2, main = "비경제
활동인구와 2차 추세모형 예측값")
legend("topleft", legend = c("pop", "fitted"), lty = 1:2)

model.matrix(reg) # 모형 계획행렬
```

```
resid=ts(resid(reg), start = c(2010, 1), frequency = 12)
ts.plot(resid(reg), ylab="residual", main = "2.2.1 잔차"); abline(h=0)
acf2(resid(reg), main="잔차의 ACF & PACF")
```

2.3.1 윈터스 가법모형

```
pop <- ts(z, start = c(2010, 1), frequency = 12)
```

Holt Winters additive model 가법형 모델

```
fit6 <- hw(pop, seasonal="additive", h = 12)
fit6$model
plot(fit6, xlab = "year", ylab = "pop", lty=1, col = "blue",
     main = "2.3.1 가법모형") #예측오차로 시계열 그림
lines(fit6$fitted, col = "red", lty = 2)
legend("topleft", lty = 1:2, col = c("blue", "red"), c("Pop", "Additive"))
ts.plot(resid(fit6), ylab="residual",
     main = "2.3.1 가법모형의 예측오차") ; abline(h= 0)
acf(resid(fit6), main = "Residual ACF")
t.test(resid(fit6), mu = 0)
```