다변량 및 시계열 데이터 분석

-단변량 시계열 데이터 분석 보고서-

2017110418 통계학과 신윤식

**1.데이터셋의 선정 및 시계열 모델의 필요성**

• 최근들어 출산율에 따른 사회 및 경제적 변화에 대한 영향이 크다는 뉴스를 자주 접하고 있습니다. 국민연금과 개인적인 미래에 대한 걱정 때문에 단변량 시계열 데이터분석 과제를 진행하면서 미래 예측을 해보고 인사이트를 얻어보고자 KOSIS인구동향조사 출생아 수 (1970~2023) 데이터 셋을 선정했고 x축이 시간과 관련된 데이터이며 출생 추세에 대한 시계열 예측을 수행할 수 있기에 ARIMA 시계열 모델을 선정했습니다.

**2. 적합한 모델의 선정**

• 먼저 정상성이 있는지 알아볼 필요가 있습니다. 차분을 통해 주기적인 패턴이나 추세를 없애 줌으로서 정상성을 확보할 수 있습니다. ADF검정을 통해 차분을 1번 하기만해도 p값이 0.011로 유의수준 0.05보다 값이 작아 정상시계열이 됨을 알 수 있으나, pmdarima 라이브러리를 통해 최적의 차분 수를 구한결과 d=3이 나와서 차분을 3번해야 함을 알 수 있었습니다.

• P와 q값을 알아보기 위해 ACF와 PACF를 계산 및 시각화 하여 최적의 p값과 q값을 알아보고자 하였습니다. ACF와 PACF 모두 신뢰구간 최초 진입시점에 따라 최적의 p와 q값이 결정되며 p=1,q=1이 결정됨을 알 수 있습니다. AutoArima를 통해 모델선정시 ARIMA(1,3,1)모델이 선정되었습니다.

라인, 그래프, 도표, 텍스트이(가) 표시된 사진

자동 생성된 설명

**3. 결과에 대한 고찰**

도표, 스크린샷, 그래프, 텍스트이(가) 표시된 사진

자동 생성된 설명• ARIMA(1,3,1)모델을 만들고 시각화를 통해 정상성을 확인해보겠습니다.

A)잔차가 평균0을중심으로 잘 퍼져 있습니다. B)히스토그램과 kde가 정규분포 모양입니다. C)QQplot에서 데이터 포인트가 직선 모양입니다. D) ACF가 첫 값으로부터 끊기는 모습으로 정상성이 있다고 말할 가능성이 높으며 이런 경우 MA도 고려해야 하는데 잘 고려하였습니다.

텍스트, 그래프, 도표, 스크린샷이(가) 표시된 사진

자동 생성된 설명텍스트, 그래프, 도표, 스크린샷이(가) 표시된 사진

자동 생성된 설명왼쪽 그림은 ARIMA모델에 대한 단순 예측이고

오른쪽그림은 one-step ahead방식입니다.

• ARIMA모델에 대한 예측을 통해 미래에도 지금과 같은 저출산율이 계속해서 이어질 수 있다는 점을 알 수 있었습니다. 단순한 예측은 주로 추세가 고려되지만 예측한 값을 이용해 추가적으로 예측하는 one-step ahead방식은 test값과 predicted값이 시각적으로 비슷하다는 것을 볼 수 있습니다. 예측모델 오차는 MAPE 값이 0.058이 나왔습니다.

•지금과 같은 추세가 이어진다면 계속해서 출산율이 감소할 것으로 예상이 되며, 이에 대한 국가적 개인적 대비가 필요하다고 생각합니다.