[시계열 평활기법]

이동평균법, 지수평활법, 홀트모형, 윈터스 모형, 분해법

[정상적 ARMA모형]

AR, MA, ARMA(AR+MA)

[비정상적 모형] : 계절성, 추세가 있는 경우

ARIMA, 계절성 ARIMA

[오차 이분산 모형] : 경제/금융 시계열

ARCH, GARCH

[다변량 시계열] : 여러 시계열 동시에 고려

벡터회귀모형(VAR) : AR의 확장

[상태공간모형]

시계열은 수평적 패턴, 추세패턴, 계절성이 포함된 패턴으로 구분할 수 있는데, 이 패턴에 따라 적절한 모형을 선택해야함.

시계열 분석 : 하나의 변수에 대한 시간에 따른 관측치

* 시계열의 특성(추세, 계절성)을 요약하고 시간에 따른 패턴 분석 / 시간에 따른 패턴을 바탕으로 모형화하고 미래값을 예측하는 것이 목적
* 시계열 패턴은 수평, 추세, 계절성이 복합된 것으로 간주함.
* 회귀분석과 달리 다른 변수를 도입하지 않고 자신의 변수의 과거패턴이 미래에도 계속된다는 가정하에 변수의 과거값을 바탕으로 미래값 예측

1. 시계열 평활기법 – 이동평균법

이동평균 (Moving Average) : 매 시점에서 직전의 N개의 데이터의 평균을 산출해 평활치로 사용함 / 시간을 x축으로 하고 추세가 그려진 그래프상 패턴을 찾기 어려우므로 smooth version line을 그렸을 때, 추세 찾기 편하므로 평활치를 사용해 line을 찾는 것

* 단순이동평균과 이중이동평균을 사용 (M으로 표현)

**단순이동평균(Simple MA)** : 시계열데이터 {X1,X2,..}가 수평적 패턴일 경우 사용, 수평적 패턴 => 수평적인 (상수함수꼴) 패턴 안에서 변동성을 가지고 추세가 나타남

t시점에서의 평활치값을 t+1시점의 값을 예측값으로 사용 : fT,1(1은 다음시점을 의미) = M T

t시점 직전의 N개의 데이터를 사용해 산술평균 내 사용(t시점의 데이터도 포함해서 N개) : N이 클수록 평활효과가 커짐(더 smooth curve) -> 따라서 N이 하나의 parameter, N이 작으면 최근의 추세를 반영하는 것 (N 사용은 주관적인 것)

**이중이동평균(Double MA)** : 시계열 데이터 {X1,X2…}가 추세 패턴을 따르는 경우 사용, 수평이 아닌 시간에 따라 증가하는 추세(선형함수꼴)와 같은 추세 패턴 하에서 변동성을 가지고 그래프가 나타날 때

XT = c + b\*t + at 일 때, 단순이동평균의 기댓값 = c + b\*t – (N-1)/2 \*b 즉, c + b\*t가 기댓값이 아니라 추가적인 term이 빼짐(gap이 생기는 것). => 단순이동평균값으로는 추세를 늦게 따라간다고 볼 수 있음

이 gap을 보정하기 위해 이중 이동평균이라고 함 -> 단순이동평균에 대해 이동평균을 한번 더 취해준 꼴 -> 즉, 이중이동평균은 각각의 t-N+1~t시점까지의 값으로 t시점의 단순이동평균을 구하고, t-N+1 ~ t시점까지 각각의 단순이동평균들의 산술평균을 구해주는 것(이중이동평균)을 구해줌 -> 이 둘간의 차(단순이동평균-이중이동평균) \* (2/(N-1))을 통해 기울기 추정가능 (기댓값을 통해 증명가능)

* 선형추세는 parameter가 2개(절편과 기울기) : 하나의 통계량으로 두개 추정이 안되기 때문

T시점까지의 데이터를 통해 T+1을 어떻게 예측하는가 ?

fT,1 = E(XT+1 | XT , XT-1 , …) = c + b(T+1) ; 조건부 기댓값이며, c와 b는 앞에서 사용한 추정치 사용할 것

f hat T,1 = c hat + b hat\*(T+1) = 2\*단순이동평균 – 이중이동평균 + bhat

k단계 이후의 예측치

fT,k = E(XT+k | XT , XT-1 , …) = c + b(T+k), 각 c,b를 추정치 사용해 예측값 구함 ( 2\*단순이동평균 – 이중이동평균) + k\*bhat

[예측성능 척도]

* 예측 오차 : 주로 특정시점에서 다음 시점을 예측하고, 다음 시점의 실제값과 비교해 예측 오차를 산출 ex) et,1 = X t+1 - f t,1 (실제값-예측값) : t시점에서 한단계 다음의 예측오차임
* n개의 시점에서 예측 오차 산출 : MSE(평균제곱오차 : sigma(et^2)/n) – 제곱이기 때문에 원래의 단위로 환산 불가 -> Root씌워 사용하면 원래단위랑 맞아짐 / RMSE : root(MSE) / MAD(평균절대오차 : sigma(|et|)/n) / MAPE(평균절대퍼센트오차 : sigma(|et,1/Xt+1 |) \* 100/n : 퍼센트로는 크고 작음을 비교하기 수월하므로 / 상대오차상 사용하는 것 / 절댓값(실제값과 예측값간 차이)/실제값 이 상대오차

1. 시계열 평활 기법 – 지수평활법

지수 평활법 : 평활치를 구하는데 전체 데이터를 사용하되, 시간에 따라 다른 가중치, 과거일수록 지수적으로 감소하는 가중치 사용

* 단순 지수평활, 이중 지수평활, 홀트 모형이 있음

**단순 지수평활(Simple Exponential Smoothing)** : 시계열 데이터 {X1,X2,…}가 수평적 패턴

t시점에서의 지수평활치 (S사용) : St =a\*Xt + a(1-a)\*Xt-1 + a((1-a)^2)\*Xt-2 + …

t+1 : S(t+1) = a\*X(t+1) + (1-a)\*S(t) : 평활상수인 a (0<alpha<1)가 작을수록 평활효과가 큼(변동을 줄이는 효과가 큼) : a가 작음은 최근치의 가중치가 작은 것 => 전체평균에 가까워짐, 그래서 주로 alpha 는 *작은 값을 사용 (평활이 주 목적*이므로 , 0.1,0.2,0.3…)

예측 : f (t,1) = S(t)로 X(t+1)을 예측함

**이중 지수평활(Double ES)** : 시계열 데이터 {X1,X2,…}가 추세 패턴

X(t) = c + b+t + a(t) : 2개의 파라미터 즉, 단순 지수평활치의 기댓값과 시계열 기댓값(X에 대한 기댓값)간 격차가 존재함

단순 지수평활치의 기댓값 = c + b\*t – (1-a)\*b/a (추가적인 term이 빼지는 것을 확인가능)

따라서 이중 지수평활로 보정할 것

S(t) = a\*X(t) + (1-a)\*S(t-1) : 단순 / S2(t) = a\*S(t) + (1-a)\*S2(t-1) : 이중 (즉, S2(t)는 직전의 이중 지수평활치에 최근의 단순지수평활치를 가중평균한 값)=> E(St)-E(S2t) = (1-a)\*b/a

이 두 통계량을 통해 c,b 추정가능 (bhat = a/(1-a) \* (S(T)-S2(T))

*질문 : S2(t-1) 등은 어떻게 구하는가 ? : n이 크다면 상수항은 0에 수렴하므로 무시, n이 크지 않다면 초기 평활값을 0시점에서 선형 추세 모형의 최소제곱추정량으로 구함*

*출처 :* <https://datalabbit.tistory.com/76>

-예측

f(T,1) = E(X(T+1)|X(T),X(T-1),…) = c + b(T-1)이므로

fhat -> 각 b,c의 hat값 사용 = 2\*단순지수평활치 – 이중지수평활치 + bhat

k단계 이후 예측치는 앞에서 1자리에 k 대입해주면 됨

**홀트 모형(Holt’s Model)** : 시계열 데이터 {X1,X2,…}가 추세 패턴

수평수준과 추세를 각각 갱신하는 모형임 (수평수준 : Lt = a\*Xt + (1-a)\*(L(t-1)+b(t-1) / 추세 : b(t) = Beta \* (L(t)-L(t-1)) + (1-beta)\*b(t-1), alpha, beta모두 0~1사이 값

T시점에서의 T+k 시점값 예측 f(T,k) = L(T)+k\*b(T)

반드시 이중지수평활치보다 홀트모형이 항상 좋은 것은 아님

[계절성 고려 모형]

* 추세와 계절성이 있는 시계열에 적용, 윈터스 모형과 분해법

1. **윈터스 모형** : 홀트 모형에 계절성(1,2,3,4분기 / 주기성있게 변화하는)을 추가 반영(계절성지수)하여 확장 / 가법(additive)모형과 승법(multiplicative)모형이 있음

[승법 모형]

계절성지수 : s(t) : t=1,2,…,m (m: 분기별이면 4)

홀트 모형처럼 수평수준, 추세, 계절성을 각각 갱신함

수평수준 L(t) : a\*x(t)/s(t-m) + (1-a)(L(t-1)+b(t-1))

추세 b(t) = beta(L(t)-L(t-1)) + (1-beta)\*b(t-1)

계절성 s(t) = gamma \*x(t)/L(t) + (1-gamma)\*s(t-m) : s(t-m)은 이전 계절성을 의미

각 평활상수 alpha, beta, gamma는 0~1 , 초기치가 필요하며(데이터가 작을 때 특히 중요) 계절성지수는 평균이 1이 되도록 조정해야함(승법에서)

시점 T+k의 값 예측 f(T,k) = (L(T)+k\*b(T))\*S(T-m+k) ; 승법이기 때문에 곱한 꼴

1. **분해법** : 추세와 계절성을 분해한 후 예측할 때 다시 결합 / 가법모형과 승법 모형이 있음

텍스트, 전자제품, 스크린샷, 소프트웨어이(가) 표시된 사진

자동 생성된 설명(합이 m이므로 평균은 1이라는 제약조건)

분해법에 따른 예측절차

1. 중심 이동평균으로 평활치 산출

Centered moving avg : t시점의 평활치로, 양쪽 데이터, 즉 전후 데이터 사용해 중심의 데이터를 예측

주기 홀수이면, 중앙(모두 가중치 1), 주기가 짝수라면 m=10이라면, 양쪽끝값 0.5가중치, 나머지는 가중치 1가지고 평활치 구함

1. 추세제거(detrended) 시계열 산출 -> 계절성 지수 구하기 위함

텍스트, 전자제품, 스크린샷, 소프트웨어이(가) 표시된 사진

자동 생성된 설명

1. 계절성 지수 산출

계절별로, 추세 제거 시계열값의 평균으로 계절성 지수를 산출하되, 계절성 지수 평균이 1이 되도록 보정해줌(승법에서의 제약조건 만족을 위함)

1. 계절성 제거(deseasonalized) 시계열 산출

위 그림의 DX(t)(S)값보면 됨. 실제값에 계절성지수 나눠준 형태

1. 회귀모형으로 추세 추정 (계절성지수 제거한 것으로 추세를 결론적으로 추정하는 것)

텍스트, 전자제품, 스크린샷, 소프트웨어이(가) 표시된 사진

자동 생성된 설명

계절성 제거 시, 추세 파악이 쉬워짐 -> 좀 더 smooth한 형태로 그래프가 그려짐

R squared 등으로 비교해 적절한 회귀식 도출

1. 추세 및 계절성지수를 결합해 예측치 산출

* 계절별 계절성을 볼 때에도 월별 데이터가 있다면 계절주기는 월 기준, 즉 12를 사용해야함.