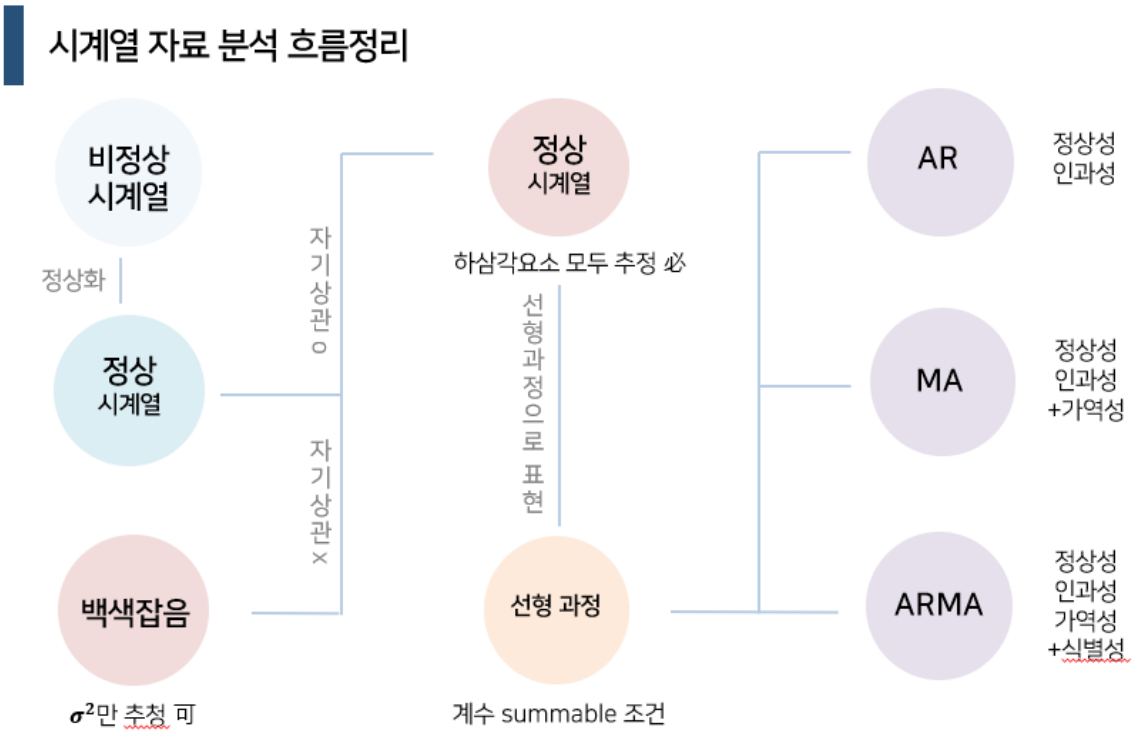
시계열자료분석팀 3주차

[목차]

1. 지난 시간 복습
2. ARIMA
   2.1 ARIMA 정의
   2.2 ARIMA 적합 절차
3. SARIMA
   3.1 SARIMA 정의
   3.2 SARIMA 적합 절차
4. 이분산모형
   4.1 수익률과 조건부 이분산성
   4.2 ARCH
   4.3 GARCH
5. ARMAX
6. VAR
7. 시계열과 ML

1. 지난 시간 복습



[동기님 작품^_^]

- ACF 와 PACF 를 통해 모형 선정

	AR(p)	MA(q)	ARMA(p,q)
ACF	지수적으로 감소	q+1차부터 절단	지수적으로 감소
PACF	p+1차부터 절단	지수적으로 감소	지수적으로 감소

2 주차에 배운 모형은 모두 정상화를 거친 데이터에 적용하는 모형들이었습니다. 오늘은 정상화를 진행하지 않은 원본 데이터에 적용할 수 있는 비정상 시계열 모형을 배워보겠습니다!

2. ARIMA 모형

2.1 ARIMA 모형의 정의

첫 번째로 배울 비정상 시계열 모형은 ARIMA 입니다. ARIMA 는 이름에서 확인할 수 있는 것처럼 차분과 ARMA 모형이 결합된 모형입니다. d 차 차분 결과, 오차 Y_t가 ARMA(p,q) 모형을 만족하는 원본 시계열 데이터를 **자기회귀누적이동평균과정 ARIMA(p,d,q)**를 따른다고 정의합니다. (여기서 p 는 AR 의 차수, d 는 차분의 차수, q 는 MA 의 차수를 의미합니다!) 식으로는 다음과 같이 표현할 수 있습니다.

$$\phi(B)(1-B)^d X_t = \theta(B)Z_t$$

$$\left(1-\phi_1 B-\phi_2 B^2-\cdots-\phi_p B^p\right)(1-B)^d X_t = \left(1+\theta_1 B+\theta_2 B^2+\cdots+\theta_q B^q\right)Z_t$$

(위 식에서 $\phi(B)$는 AR 의 특성방정식, $\theta(B)$는 MA 의 특성방정식,

$(1-B)^d$는 d 차 차분을 의미해 ARIMA 모형의 정의를 다시 한번 확인할 수있습니다!)

위 내용을 정리하자면, ARIMA 모형은 데이터가 **polynomial trend 를 가지고 있으며, 오차가 ARMA(p,q)를 따른다고 판단될 때**만 사용할 수 있습니다. ARIMA 모형의 장점은 2 주차에 배운 모형들과 달리 차분을 포함하고 있어 간결하게 표현할 수 있다는 것과 여전히 선형과정을 따르기 때문에 예측이 용이하다는 것입니다.

cf) ARIMA 는 차분과 ARMA 를 결합한 모형인데 왜 ARDMA 가 아닌 ARIMA 일까요?!

ARIMA 에서 I 는 누적(Integration)을 의미합니다. 아래 식을 통해 확인해보겠습니다!

$$\phi(B)(1-B)X_t = \theta(B)Z_t$$

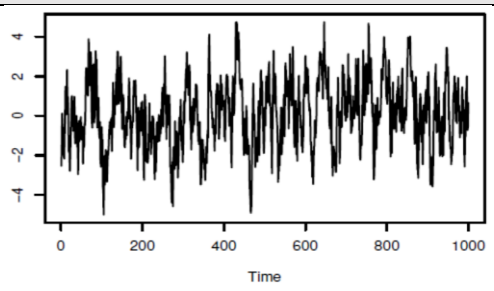
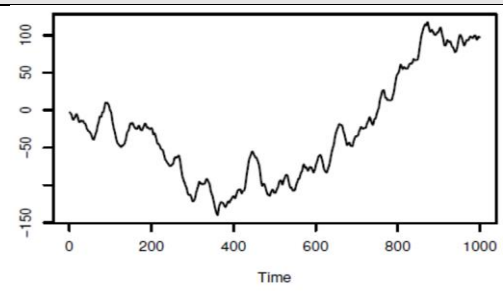
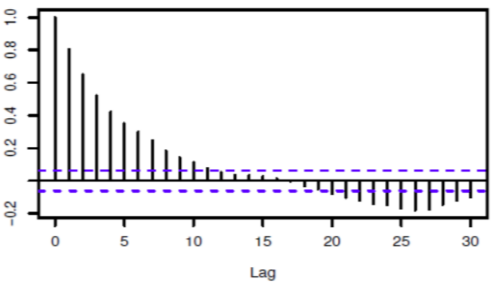
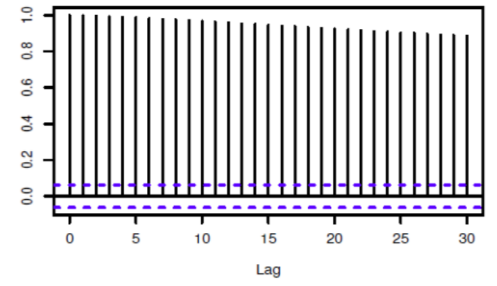
$$Y_t = (1-B)X_t = X_t - X_{t-1}$$

$$X_t = X_{t-1} + Y_t = (X_{t-2} + Y_{t-1}) + Y_t = \cdots = X_0 + \sum_{j=1}^{t} Y_j$$

위 식과 같이 X_t는 $\boldsymbol{Y_j}$**의 누적합**으로 볼 수 있으며, 이는 랜덤워크의 확장으로 해석할 수 있습니다. 따라서 ARIMA 모형을 자기회귀**누적**이동평균모형이라고 부릅니다.

2.2 ARIMA 모형의 적합 절차

[1] TS plot 과 ACF 그래프를 통해 정상/비정상 시계열 여부를 판단합니다.

	정상 시계열	비정상 시계열
TS plot		
ACF		

위 그림처럼 정상 시계열은 ACF 가 지수적으로 감소하지만 비정상 시계열은 ACF 가 천천히 감소하는 감소하는 것을 기억해 적절한 모형을 선택할 수 있습니다.

[2] 비정상 시계열에서 추세가 관측된다면, 차분을 통해 정상화를 진행합니다. 이때 과대차분이 되지 않도록 주의해야 합니다!

과대차분이란 필요 이상으로 차분을 과하게 진행하는 것입니다. 정상화가 완료되었음에도 추가적으로 차분을 진행할 경우 정상성에는 문제가 없지만, ACF 가 복잡해지거나 분산이 커지며, 불필요한 상관관계를 생성해 적합 과정이 복잡해질 수 있습니다. 따라서 적절한 차분의 차수를 결정해 과대차분을 방지해야 합니다. 일반적으로 대부분의 데이터는 1,2 차 차분만으로 정상성을 만족합니다.

[3] 모형 적합 절차에 따라 p, q 의 차수를 결정하고, 모수를 추정하고 진단까지 마친 모형을 통해 예측을 진행합니다.

지금까지 다항 **추세가 존재하는 비정상 데이터**에 ARMA 를 적용하는 ARIMA 모형에 대해 배웠습니다. 그렇다면 **추세와 계절성이 모두 존재하는 경우**에는 어떤 모형을 사용할까요?

3. SARIMA 모형

두 번째로 배울 비정상 시계열 모형은 SARIMA 입니다.

1 주차에 배운 정상화에서의 계절성은 결정적(deterministic) 계절성으로, **모든 주기에서 계절성이 동일함**을 가정했습니다. 하지만 현실에서는 모든 계절성분이 결정적이지는 않으며, 다른 성분들과의 상관관계가 존재할 수 있습니다. 이렇게 주기가 변함에 따라 계절성도 변화할 수 있음을 반영하여 **확률적 분석**을 하는 모형이 SARIMA 모형입니다.

3.1 SARIMA 모형의 정의

SARIMA 모형은 Seasonal ARIMA 의 줄임말로, **추세와 계절성이 모두 존재하는 비정상 데이터**에 적용할 수 있는 모형입니다. 가장 확장된 형태의 ARMA 모형으로, 계절성 사이의 상관관계에 대해 모델링하는 확률적 접근법으로 이해하시면 됩니다!

주기가 12 인(s=12) 예시 데이터를 가지고 SARIMA 모형이 계절성을 어떻게 다루는지 알아보도록 하겠습니다!

	January	February	...	December
Year 1	Y_1	Y_2	...	Y_{12}
Year 2	Y_{13}	Y_{14}	...	Y_{24}
⋮	⋮	⋮	⋮	⋮
Year r	$Y_{1+12(r-1)}$	$Y_{2+12(r-1)}$	...	$Y_{12+12(r-1)}$

[Step1]

위 데이터에서 **각 열**을 계절성분으로 정의하겠습니다. 각 열을 고정한 후, 1 월부터 12 월까지 모든 열은 동일한 ARMA(P, Q) 모형을 따른다고 가정합니다. (이때 같은 열에 속한 Y_j끼리는 상관관계가 존재합니다.) 위 내용을 식으로 표현하면 아래와 같습니다.

$$Y_{j+12t} - \Phi_1 Y_{j+12(t-1)} - \cdots - \Phi_P Y_{j+12(t-P)} = U_{j+12t} + \Theta_1 U_{j+12(t-1)} + \cdots + \Theta_Q U_{j+12(t-Q)}$$

$$(where\ t = 0,1, \dots, r\ \ U_t \sim WN(0, \sigma_U^2))$$

$$\updownarrow$$

$$\Phi(B^{12}) Y_t = \Theta(B^{12}) U_t$$

(j ∈ [1,12]는 month 에 해당하며, j가 달라져도 Φ와 Θ는 동일하다고 가정합니다.)

[Step2]

지금까지는 같은 month 에 속하는 시계열끼리의 상관관계를 모델링하는 과정이었습니다. 이번에는 열이 아닌 **행을 고정**해서 **각 월끼리의 상관관계**에 대해 살펴보겠습니다. 위 식에서 Y_t는 U_t에 의해 표현되고 있기 때문에 Y_t와 Y_j의 correlation 은 U_t와 U_j에 대해 모델링을 진행해 구할 수 있습니다.

U_t 를 ARMA(p, q)를 따르는 시계열이라고 생각해보겠습니다. U_t 를 백색잡음이 아닌 ARMA(p,q)를 따르는 시계열이라고 가정하는 이유는 $\{Y_1, \dots, Y_{12}\}$ 와 같이 행에도 correlation 이 존재할 수 있음을 반영하기 위함입니다. 위 과정을 식으로 나타내면 아래와 같습니다.

$$\phi(B)U_t = \theta(B)Z_t, \quad Z_t \sim WN(0, \sigma^2)$$

[Step3]

Step1 과 Step2 과정을 합쳐서 하나의 식으로 정리하면 다음과 같습니다.

$$\Phi(B^{12})Y_t = \Theta(B^{12})\phi^{-1}(B)\theta(B)Z_t$$

$$\phi(\mathrm{B})\Phi(\mathrm{B}^{12})\mathrm{Y_t} = \theta(B)\Theta(B^{12})Z_t, \quad Z_t \sim WN(0, \sigma^2)$$

여기까지의 과정이 Seasonal ARMA 로 SARMA(p, q)X(P, Q)입니다.

[Step4]

SARMA 에 차분을 더해서 SARIMA 모형식을 완성합니다. **SARIMA(p,d,q)X(P,D,Q)** 모형은 아래 식을 통해 표현할 수 있습니다.

$$\phi(B)\Phi(B^{12})(1-B)^d(1-B^{12})^D X_t = \theta(B)\Theta(B^{12})Z_t, \quad (Z_t \sim WN(0, \sigma^2))$$

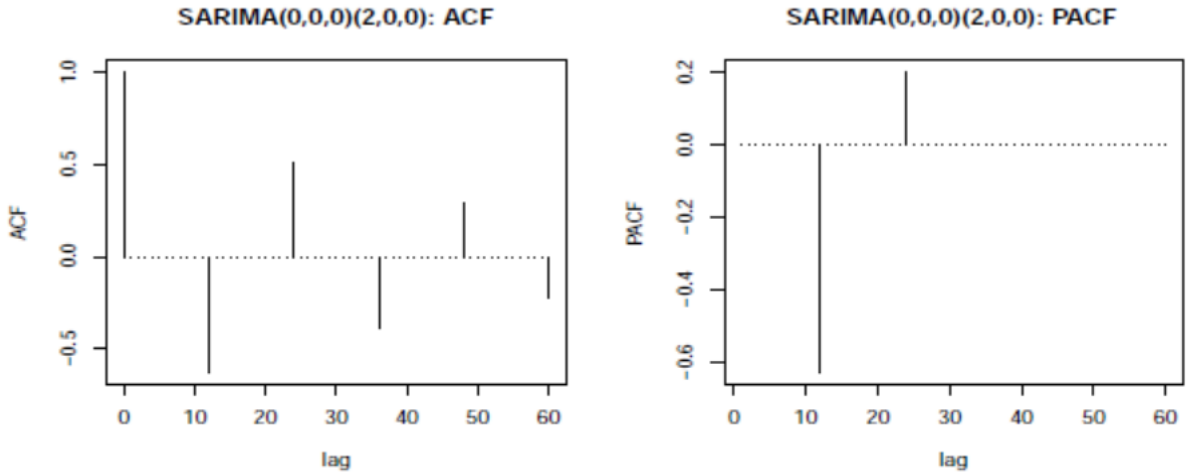
소문자 (p,d,q)는 전체 시계열에 대한 차수를 의미하며, 대문자 (P,D,Q)는 주기 패턴에 대한 차수를 의미함을 주의해야 합니다! 따라서 $(1-B)^d$는 전체 시계열의 추세에 대해 d 차 차분을 적용한 것이고, $(1-B^{12})^D$ 는 계절성분이 갖는 D 차 추세에 대해 lag-12 차분을 D 번 적용한 것입니다.

3.2 SARIMA 적합 절차

[1] TS plot 과 잔차를 확인하여 이분산성이 나타나는 경우 분산 안정화를 진행합니다.

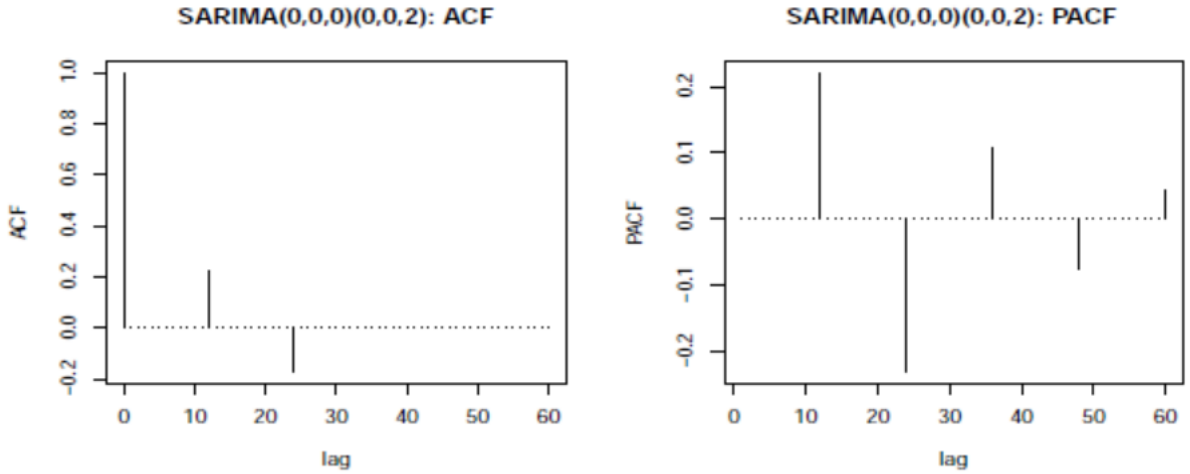
[2] d 차 차분 또는 lag-d 차분 진행 여부를 결정합니다. ARIMA 와 마찬가지로 ACCF 그래프가 느리게 감소한다면 차분이 필요하다고 판단할 수 있으며, 느리게 감소하는 동시에 규칙적으로 구불거리는 형태가 보인다면 계절 차분을 진행해야 함을 의미합니다. 이때 과대차분이 되지 않도록 주의해야 합니다!

[3] P, Q 와 p, q 의 차수를 결정합니다. p 와 q 는 ARMA 와 동일하게 ACF, PACF 를 통해 구할 수 있습니다. P 와 Q 는 계절성분의 모수이기 때문에 그래프를 해석하는 방법에 차이가 존재합니다. 아래 예시를 통해 확인해보겠습니다!



위 그림은 SARIMA(0,0,0)(2,0,0)을 따르며 주기는 12 인 데이터의 ACF 와 PACF 입니다. SARIMA(0,0,0)(2,0,0)이란 계절성분, 즉 고정된 열들이 ARMA(2,0) (=AR(2)) 모형을 따르는 것입니다. 클린업 2 주차에 AR 모형의 ACF 는 지수적으로 감소, PACF 는 시차 이후 절단됨을 배웠습니다. 두 그래프 모두 시차의 배수에만 값이 존재하며, PACF 는 lag 24 이후로 절단됨을 확인할 수 있습니다.

이처럼 SARIMA 의 ACF 와 PACF 는 주기의 배수 간격으로 그래프를 해석해야 합니다! 아래 예시를 통해 한 번 더 확인해보겠습니다.



위 그림은 마찬가지로 주기가 12 인 SARIMA(0,0,0)(0,0,2)의 ACF 와 PACF 입니다. 위 예시와 동일하게 시차 간격으로 해석했을 때 MA(2)의 ACF 와 PACF 의 특징을 만족함을 알 수 있습니다.

만약 ACF 와 PACF 모두에서 절단되는 부분을 찾을 수 없다면 ARMA 모형을 적합해야 하고, 2 주차에 배운 것처럼 IC 가 가장 높은 모수의 차수를 선택해야 합니다.

[4] 모수를 추정한 후 예측을 진행합니다!

지금까지 비정상 데이터에 적용하는 비정상 시계열 모형에 대해 알아봤습니다. 비정상 시계열 모형은 앞에서 언급한 것과 같이 정상화 과정을 따로 진행하지 않아도 된다는 장점이 있습니다. 따라서 예측 결과 역시 원본 데이터에 대한 최종 예측값으로, 추세와 계절성을 따로 더하지 않아도 됩니다! 예~

4. 이분산 시계열 모형

지금까지 우리가 배운 시계열 모형들은 시간에 따른 분산의 변화가 없다고 가정한 후 평균의 움직임에 관심을 갖는 모형들이었습니다. 하지만 수익률, 주가, 환율 등 금융관련 시계열에서는 분산이 과거 자료에 의존하는 특성을 가정하고 있습니다. 즉 **시간에 따른 이분산성**에 관심을 가지고 분석을 진행하는 것입니다. 경제학 분야에서 위험을 측정할 때 **변동성(volatility)**를 주로 사용하며, 이는 통계학에서의 **조건부 분산(conditional variance)**로 표현할 수 있습니다. 즉, 지금부터 배울 이분산 시계열 모형은 조건부 분산을 시간의 함수로 표현하는 시계열 모형이라고 생각하시면 됩니다!

cf) 변동성이란 주식시장 등 자산시장에서 상품의 가격이 변동하는 정도!

4.1 수익률과 조건부 이분산성

이분산 시계열 모형인 ARCH 와 GARCH 에 대해 배우기 위해 수익률과 조건부 이분산성에 대해 알아보도록 하겠습니다.

1) 수익률

수익률에 대한 대표적인 정의는 simple return 과 log return 이 있습니다.

- Simple return : $R_t = \frac{P_t - P_{t-1}}{P_{t-1}}$
- Log return : $r_t = log\, P_t - log\, P_{t-1} = log\,(1 + R_t) \approx R_t$

Log return 은 덧셈으로 표현되며 log 를 통해 분산을 안정화시키는 효과도 있기 때문에 앞으로 나오는 모든 수익률은 log return 을 의미합니다!

2) 조건부 이분산성(Conditional Heteroskedasticity)

조건부 이분산성과 반대되는 개념인 **조건부 등분산성**에 대해 먼저 알아보겠습니다. 조건부 등분산성이란 구조적인 변동성이 이전 기간의 변동성의 영향을 받지 않는 것이며, 아래 식을 통해 표현합니다.

$$\mathrm{Var}(r_t|\mathcal{F}_{t-1}) = \text{constant}$$

$$\text{where } \mathcal{F}_{t-1} \text{ is the } \sigma - \text{field generated by historical information}$$

이를 통해 조건부 이분산성의 식을 표현하면 아래와 같습니다.

$$\mathrm{Var}(r_t|\mathcal{F}_{t-1}) \neq \text{constant}$$

정리하자면 **조건부 이분산성**이란 변동성이 시점에 의존하는 것, 미래의 변동이 현재까지의 상황에 의존하는 것으로 정의할 수 있습니다. 많은 금융 시계열 자료는 이러한 조건부 이분산성을 가지기 때문에 **수익률에 대해 모델링** 하는 모형들에 대해 배워보겠습니다!

4.2 ARCH(Auto- Regressive Conditional Heteroscedasticity)

자기회귀이분산모형 ARCH 는 t 시점의 오차 변동성인 **조건부 분산** σ_t^2를 과거시점들의 오차항으로 설명하는 모형입니다. ARCH(1)은 아래 식으로 나타낼 수 있습니다!

$$r_t = \sigma_t \varepsilon_t$$

$$\varepsilon_t \sim iid\ N(0,1)$$

$$r_t{}^2 = (a_0 + a_1 r_{t-1}{}^2)\varepsilon_t^2 = \sigma_t^2 \varepsilon_t^2$$

$$\sigma_t{}^2 = Var(r_t | F_{t-1}) = a_0 + a_1 r_{t-1}{}^2$$

(위 식에서 F_{t-1}은 t-1 시점까지 모든 정보의 집합을 의미하며,

$r_t | F_{t-1}$는 t-1 까지의 모든 정보를 다 알고 있을 때의 t 시점의 수익율입니다.)

위 식에서 $\sigma_t{}^2$은 r_{t-1}^2에 의존하고 있기 때문에 조건부 이분산임을 알 수 있습니다.

ARCH(m)은 $\sigma_t{}^2$에 AR(m) 구조를 가정한 것으로, 아래 식을 통해 표현합니다.

$$r_t = \sigma_t \varepsilon_t$$

$$\sigma_t{}^2 = a_0 + a_1 r_{t-1}{}^2 + \cdots + a_m r_{t-m}{}^2 \approx r_t{}^2$$

ARCH(m)의 가장 큰 특징은 **비선형 모형**이라는 것인데, ARCH(1)을 통해 알아보겠습니다.

$$r_t^2 = \sigma_t^2 \varepsilon_t^2 = (\alpha_0 + \alpha_1 r_{t-1}^2)\varepsilon_t^2$$

$$= \alpha_0 \varepsilon_t^2 + \alpha_1 \varepsilon_t^2 \{(\alpha_0 + \alpha_1 r_{t-2}^2)\varepsilon_{t-1}^2\}$$

$$= \alpha_0 \varepsilon_t^2 + \alpha_0 \alpha_1 \varepsilon_t^2 \varepsilon_{t-1}^2 + \alpha_1^2 r_{t-2}^2 \varepsilon_t^2 \varepsilon_{t-1}^2$$

$$\vdots$$

$$= \alpha_0 \sum_{j=0}^{n} (\alpha_1^j \varepsilon_t^2 \varepsilon_{t-1}^2 \cdots \varepsilon_{t-j}^2) + \alpha_1^{n+1} r_{t-n-1}^2 \varepsilon_t^2 \varepsilon_{t-1}^2 \cdots \varepsilon_{t-j}^2$$

위 식과 같이 곱셈식으로 표현되었기 때문에 비선형적 모델임을 기억해야 합니다! 추가적으로,

α_1이 0 과 1 사이에 있을 때 정상성 역시 만족함을 알 수 있습니다!

4.3 GARCH(Generalized Auto-Regressive Conditional Heteroscedasticity)

ARCH 모형은 m 의 값이 커지면 추정해야 할 모수가 많아지고, 추정량의 정확도가 떨어진다는 단점이 존재합니다. 따라서 이를 해결하기 위해 제시된 일반화된 모델이 **GARCH 모형으로 일반화자기이분산모형**입니다. ARCH 모형은 σ_t^2에 AR 모형을 가정했던 것처럼, GARCH 모형은 σ_t^2에 ARMA 모형을 가정한 것으로 ARCH 보다 확장된 모형으로 생각하시면 됩니다!

r_t^2은 σ_t^2에 근사하고, 둘 사이의 오차를 $\eta_t = r_t^2 - \sigma_t^2$ 라고 할 때, GARCH(1, 1) 모형은 다음과 같이 표현할 수 있습니다.

$$r_t = \sigma_t \varepsilon_t \sim N(0, \sigma_t^2)$$

$$r_t^2 = \alpha_0 + (\alpha_1 + \beta_1) r_{t-1}^2 + \eta_t - \beta_1 \eta_{t-1}$$

$$r_t^2 = \alpha_0 + \alpha_1 r_{t-1}^2 + \beta_1 r_{t-1}^2 + r_t^2 - \sigma_t^2 - \beta_1 (r_{t-1}^2 - \sigma_{t-1}^2)$$

$$\sigma_t^2 = \alpha_0 + \alpha_1 r_{t-1}^2 + \beta_1 \sigma_{t-1}^2$$

주황색으로 표시된 부분이 AR(1) 부분이며, 파란색 부분이 MA(1)을 나타냅니다!

이를 일반화하면 GARCH(p, q) 모형은 아래와 같이 표현할 수 있습니다.

$$r_t = \sigma_t \varepsilon_t \sim N(0, \sigma_t^2)$$

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^{p} \alpha_i r_{t-i}^2 + \sum_{j=1}^{q} \beta_i \sigma_{t-j}^2$$

5. ARMAX

ARMAX 모형은 ARMA 모형에 독립변수로 외부요인(e**X**ogenous)을 추가한 모형입니다. 이때 추가된 독립변수는 연속형일 수도, 범주형일 수도 있습니다. 일반적으로 아래 식과 같이 표현하며, Y_t와 X_t의 관측값 수는 동일해야 합니다.

$$\phi(B)Y_t = \theta(B)Z_t + \beta_0 + \beta_1 X_t$$

Y_t를 예측하고자 하는 주가, X_t를 날씨 데이터로 예를 들면, 날씨를 반영해 주가를 예측하고자 할 때 사용할 수 있는 모형식이 됩니다.

ARMA에 차분을 더한 모형이 ARIMA, 계절성까지 더한 것이 SARIMA였던 것과 동일하게 ARMAX 에 차분을 더하면 ARIMAX, 계절성까지 더하면 SARIMAX 모형이 됩니다. 두 모형의 식은 아래와 같습니다!

[ARIMAX] $\phi(B)(1-B)^d Y_t = \theta(B)Z_t + \beta^T \underline{X}$

[SARIMAX] $\phi(B)\Phi(B^s)(1-B)^d(1-B^s)^D Y_t = \theta(B)\Theta(B^s)Z_t + \beta^T \underline{X}$

6. VAR (Vector Auto Regressive)

VAR 모형은 이름에서 확인할 수 있는 것처럼 AR 모형에 Vector 구조를 결합한 모형입니다. 2 주차에 배운 것처럼 AR 모형은 자기 자신의 과거 관측값들로 현재의 관측값을 설명하는 모형입니다. VAR 모형은 **현재 관측값을 자기 자신의 과거 관측값들과 다른 변수의 과거 관측값으로 설명**하는 모형입니다. 식을 통해 확인해볼까요?!

$$\begin{pmatrix} X_t \\ Y_t \end{pmatrix} = \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} + \begin{pmatrix} \phi_{11} & \phi_{12} \\ \phi_{21} & \phi_{22} \end{pmatrix} \begin{pmatrix} X_{t-1} \\ Y_{t-1} \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \end{pmatrix}$$

위 식은 가장 간단한 VAR(1) 모형입니다. AR 이 일변량 자기회귀모형이었다면, VAR 은 이를 확정한 다변량 자기회귀모형이라고 생각하시면 됩니다.

$$X_t = c_1 + \phi_{11} X_{t-1} + \phi_{12} Y_{t-1} + \varepsilon_1$$
$$Y_t = c_2 + \phi_{21} X_{t-1} + \phi_{22} Y_{t-1} + \varepsilon_2$$

위와 같은 식으로도 표현할 수 있으며, 이를 통해 VAR 모형은 단순 과거 데이터만을 고려하는 것이 아니라 여러 변수들의 **의존성과 상호작용**을 고려하는 모형임을 알 수 있습니다. 이러한 이유로 VAR 은 상호연계성이 높은 경제구조를 분석할 때 용이하게 사용됩니다. VAR 의 쓰임에 대해 좀 더 자세히 알아볼까요?

1) 인과관계 분석

: 2 개 이상의 변수가 존재할 때 변수 사이에 어떤 관계가 있는지, 시차 관계가 있는지 파악하기 위한 것

2) 충격반응 분석

: 여러 변수 중 특정 변수에 예상치 못한 충격(Shock)이 발생했을 때 각 변수가 어떻게 변화하는지 분석하는 것

3) 예측오차 분산분해

: 한 변수의 변화를 설명할 때 어떤 변수가 더 중요하게 작용하는지 상대적 중요성을 찾는 것

7. 시계열과 ML

지금까지 전통적인 통계적 시계열 모형들에 대해 알아봤습니다! 시계열은 오차의 독립성 조건을 만족하지 못하기 때문에 시계열 자료를 다루는 특별한 분석법이 필요한 것이었습니다. 그렇다면 시계열 자료를 다룰 때 머신러닝을 사용할 수는 없을까요?! 정답부터 말하자면 머신러닝으로도 시계열 자료를 다룰 수 있습니다. 바로 알아볼까요?

7.1 시계열 데이터의 전처리

1) 결측치 보간

데이터 분석을 진행하다보면 다수의 결측치를 발견할 수 있고, 안정적인 분석을 위해 결측치를 보간해주어야 합니다. 하지만 예상하신 것처럼 일반적인 방법을 통해 결측치를 보간할 경우, 시계열 데이터의 특성을 반영하지 못해서 해당 시점의 평균, 분산에 왜곡이 생기는 등의 문제가 발생할 수 있습니다. 따라서 시계열 데이터는 아래의 방법을 통해 결측치를 보간합니다!

- LOCF(Last observation carried forward)
  : 직전 관측치 값으로 결측치 대체
- NOCB(Next observation carried backward)
  : 직후 관측치 값으로 결측치를 대체
- Moving Average / Moving Median
  : 직전 N 의 time window 의 평균치 / 중앙값으로 대체

일반적으로는 위 3 가지 방법을 사용하지만, 결측치를 기준으로 패턴이 급격하게 변화하는 경우에는 조금 더 복잡한 방법을 고려해야 합니다.

- 선형 보간법
  : 근사 함수가 선형(linear) 함수임을 가정

- 비선형 보간법
  : 근사 함수가 비선형(non-linear) 함수임을 가정

- 스플라인(Spline) 보간법
  : 전체 구간을 근사하는 것이 아닌, 소구간으로 분할하여 보간
  각 구간마다 함수를 적합한 후 모든 구간에서 함수가 매끄럽게 이어지도록!
  (데마팀 2 주차 클린업 참고!)

위 방법들로 보간하기 어려울 정도로 결측치가 많은 경우에는 결측치 보간 모델링을 통한 방법을 사용할 수 있습니다. 하지만 모델링을 진행하기 위해서는 충분한 데이터가 있어야 함을 기억해야 합니다!

2) 노이즈 처리(Denosing)

노이즈란 다른 외부 요인의 간섭과 같은 여러 의도하지 않은 데이터의 왜곡을 불러오는 모든 것을 의미합니다. 시계열 데이터는 시간의 흐름에 따라 변화하는 통계적 특성을 가지고 있기 때문에 노이즈 역시 같이 기록될 가능성이 높습니다. 따라서 시계열 데이터 특성에 맞는 노이즈 방법을 사용해야 합니다.

- Moving Average(MA)
  평균값으로 관측치를 대체하여 평활화 하는 방법
  노이즈가 적은 데이터에 적용할 수 있음
  (노이즈가 많은 경우 평균 역시 노이즈의 성격을 띌 수 있기 때문에!)
  금융 데이터 노이즈 제거에 자주 사용됨

- Filtering

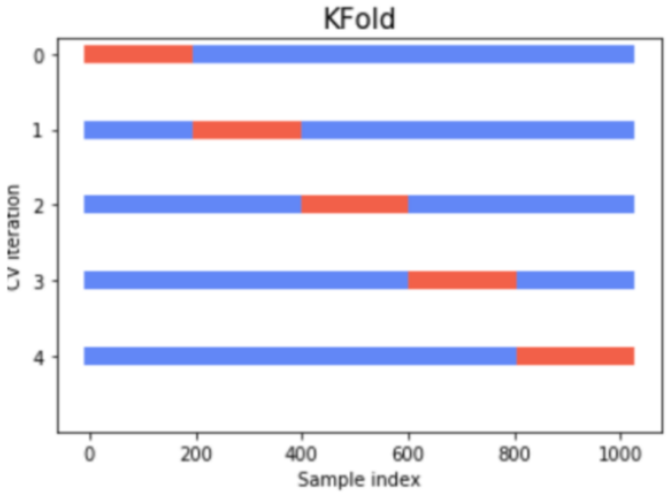
- 노이즈가 특정 분포를 따른다고 가정하고 해당 분포 값을 제거하는 방법

  ■ 가우시안 필터링 (Gaussian Filtering)
  노이즈가 정규분포를 따른다고 가정함
  중심에 가까울수록 큰 가중치를 부여하는 방법

- 칼만 필터 (Kalman Filter)
  잡음이 포함된 과거 측정값에서 현재 상태의 결합분포를 추정하는 알고리즘
  데이터의 특성에 맞는 분포를 모델링하는 것이 장점!

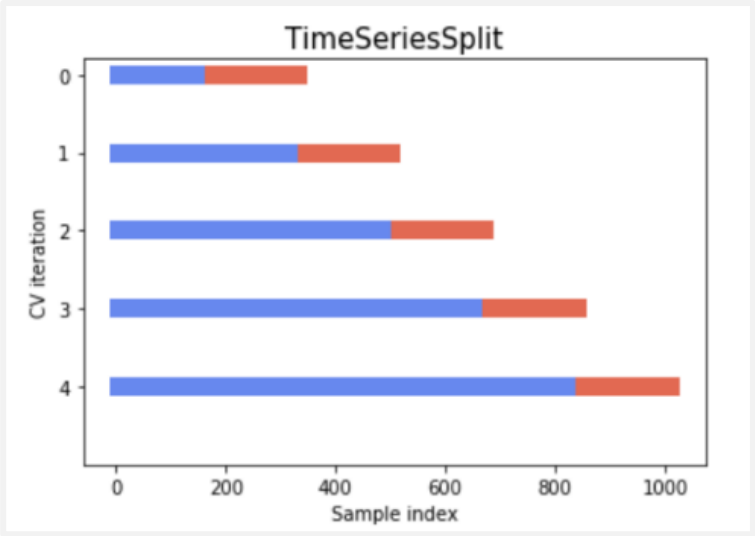
3) 시계열 데이터의 CV

교차검증(Cross Validation)이란 학습 데이터와 검증 데이터를 통해 모델의 성능을 측정하는 방법입니다. 모델의 과적합을 방지하기 위해 반드시 수행해야 하는 과정입니다.



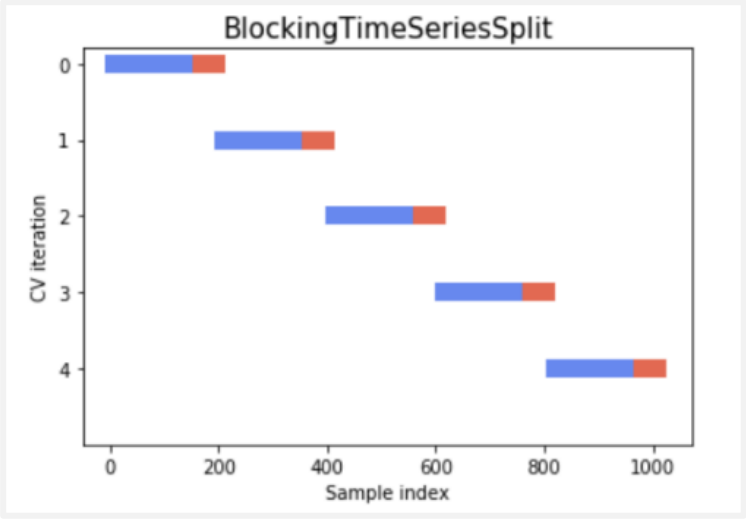
일반적으로 사용하는 K-fold CV 는 위 사진과 같이 시간 순서를 고려하지 않고 진행됩니다. 하지만 시계열 데이터는 시간에 따라 관측된 값들이기 때문에 CV 역시 이를 반영해야 합니다. 그럼 지금부터 시계열 데이터의 교차검증 방법인 Time Series CV 와 Blocked Time Series CV 에 대해 알아보겠습니다!

- **Time Series CV**



Expanding window CV 라고도 부르는데, Expanding window 란 window 를 누적하여 이동한다는 의미입니다. 이전 단계의 train set 과 validation set 을 다음 단계의 train set 으로 사용해 교차검증을 진행합니다.

- **Blocked Time Series CV**



Rolling window CV 라고도 불리는 방법입니다. Rolling window 란 동일한 사이즈의 window 를 옆으로 이동시킨다는 의미입니다. 같은 사이즈의 window 내에서 일정 비율로 train 과 validation 을 분할해 교차검증을 진행합니다.

두 방법 모두 validation 이 train 보다 미래 시점이므로, **시간의 흐름을 반영할 수 있는 교차검증** 방법입니다!

4) 클래스 불균형

모델링을 진행하다보면 클래스 불균형을 자주 만나볼 수 있습니다. 일반적인 데이터에서는 샘플링을 통해 클래스 불균형을 해결하지만, 시계열 데이터에서는 마찬가지로 시간의 흐름을 반영해야 하기 때문에 샘플링을 통해 클래스 불균형을 해결할 수 없습니다. 따라서 시계열 데이터에서는 더 적은 수를 가지는 클래스에 **가중치를 부여하는 방법**으로 클래스 불균형을 해결해야 합니다!

- Scale_pos_weight

  : 이진분류 문제에서 사용하는 파라미터로, 기본 설정은 0과 1로 라벨링 되었음을 가정한다. 수가 더 적은 쪽을 1로 설정하며, 해당 클래스에 가중치를 부여하는 방식이다.

- Class_weight

  : 마찬가지로 샘플 수가 더 적은 쪽에 가중치를 부여하는 방법

  Model.fit() 또는 model.fit_generagor()의 파라미터로 사용하여 클래스 별 가중치를 직접 제시하는 방법으로 코드를 작성

  ex) model.fit(class_weight={0:1, 1:2}) 와 같이 클래스 별 가중치 제시

- Sample_weight

  : 다중 분류에서 사용하며, 각 클래스 비율의 역수를 가중치로 계산하는 함수

  Ex) class_weight.compute_sample_weight(class_weight='balanced')

드디어 3 주의 클린업이 끝났네요! 모두 수고하셨습니다~ 클린업을 준비하면서 저조차도 '시계열 정말 머리 아파!'라는 생각을 했어서, 어떻게 해야 어렵지 않게 전달할 수 있을지에 대해 굉장히 많이 고민했던 것 같습니다. 시계열의 핵심이 무엇인지, 각각의 모형은 어떤 목적인지 흐름 위주로 설명하려고 신경 쓰면서 교안을 작성했는데 도움이 되셨기를 바랍니다 ^_^

적극적으로 질문과 대답을 주고받으며 들어주신 덕분에 저도 열정적으로 클린업을 진행할 수 있었고, 그런 여러분을 보며 '내가 더 열심히 해야겠구나!!' 생각도 많이 했네요 ㅎㅎ 주분 때는 더더 열심히 시계열 이끌어보겠습니다!!! 주분 파이팅하자구요~~ 짱시시시시시!!!!!!! 뿡