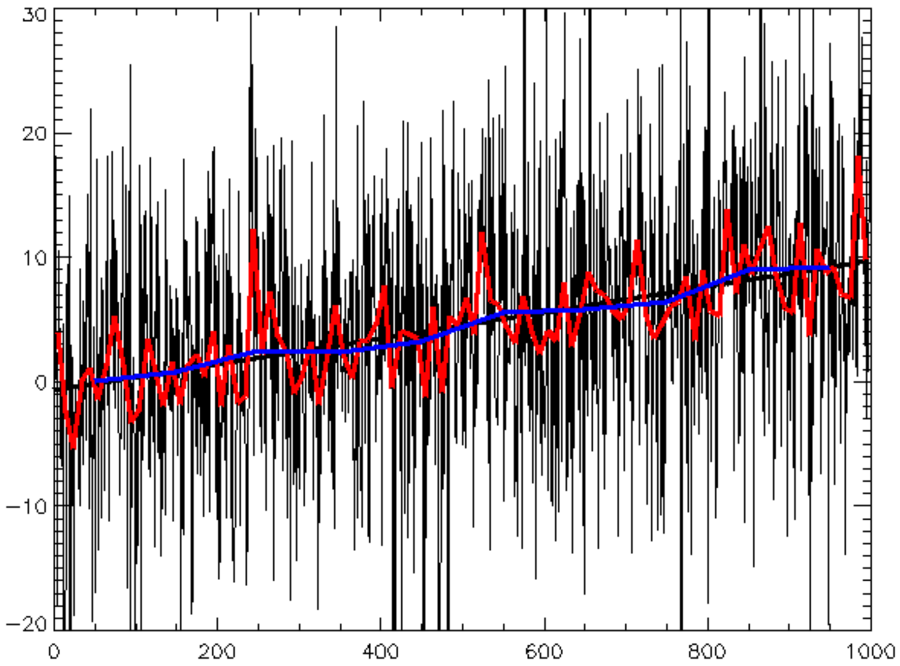
Time Series Analysis for Quant

장 연 식

시계열 분석이란 (Time series analysis)

- **Stochastic process (확률 과정)**: A sequence of infinite random variables
- **Time series (시계열):** Realization of sequences of random variables. ⇔ Observed value of stochastic process
- **Time series analysis (시계열 분석):** Statistical approach to understand the past and predict the future

- **Features of time series:**
  1. Trend (추세): Consistent directional movement
  2. Seasonal variation (계절성)
  3. Serial dependence (상관성): 과거와 현재의 상관성.
     (Ex) Volatility clustering

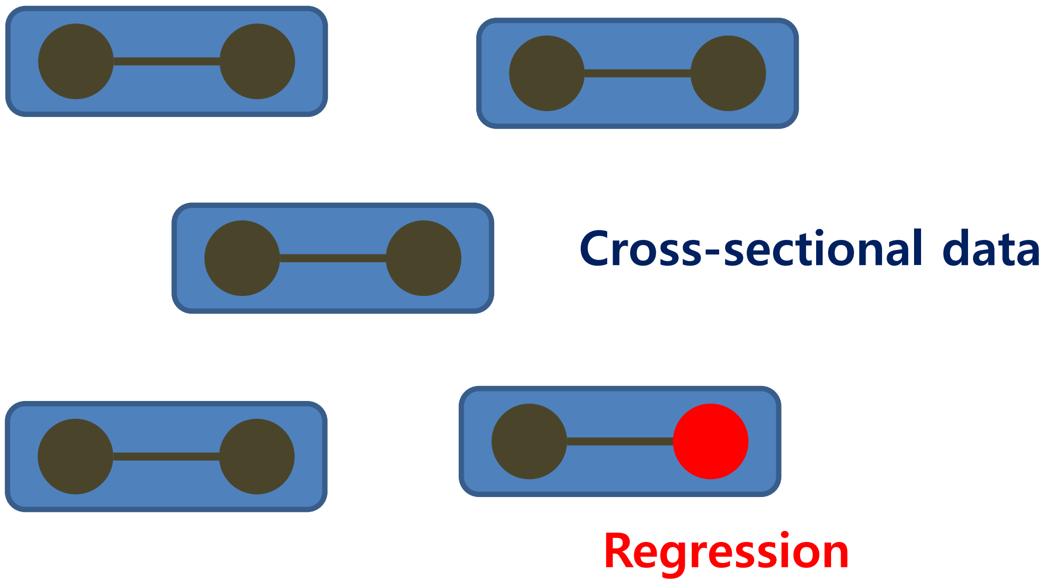


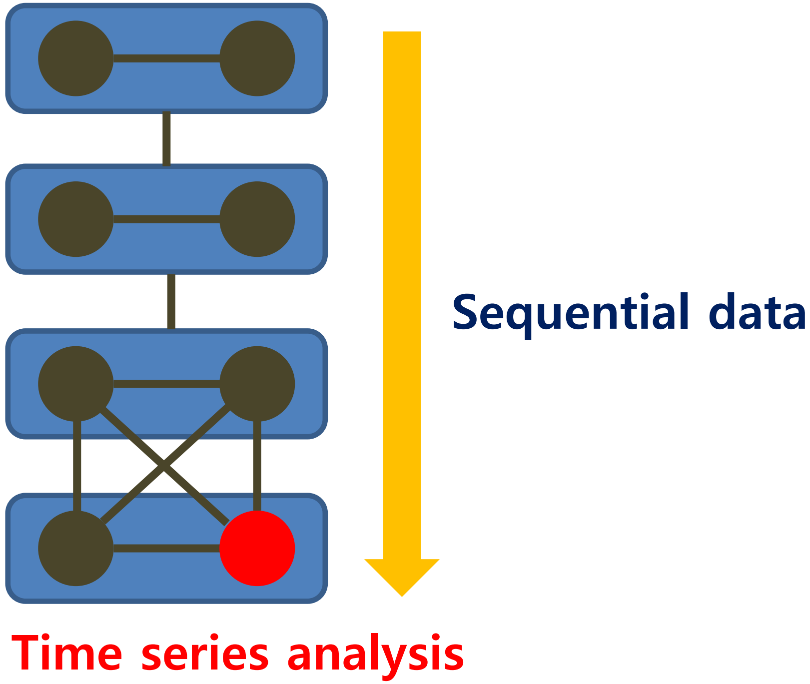
- **Applying time series in finance:**
  1. Forecasting: Predict future asset price in a statistical sense
  2. Simulate Series: Generate simulations of future scenarios
  3. Infer Relationships: Filtering, spread estimation, regime detection

시계열 분석과 횡단면 분석의 비교

시계열 데이터는 순서가 있는 sequential 데이터이다.

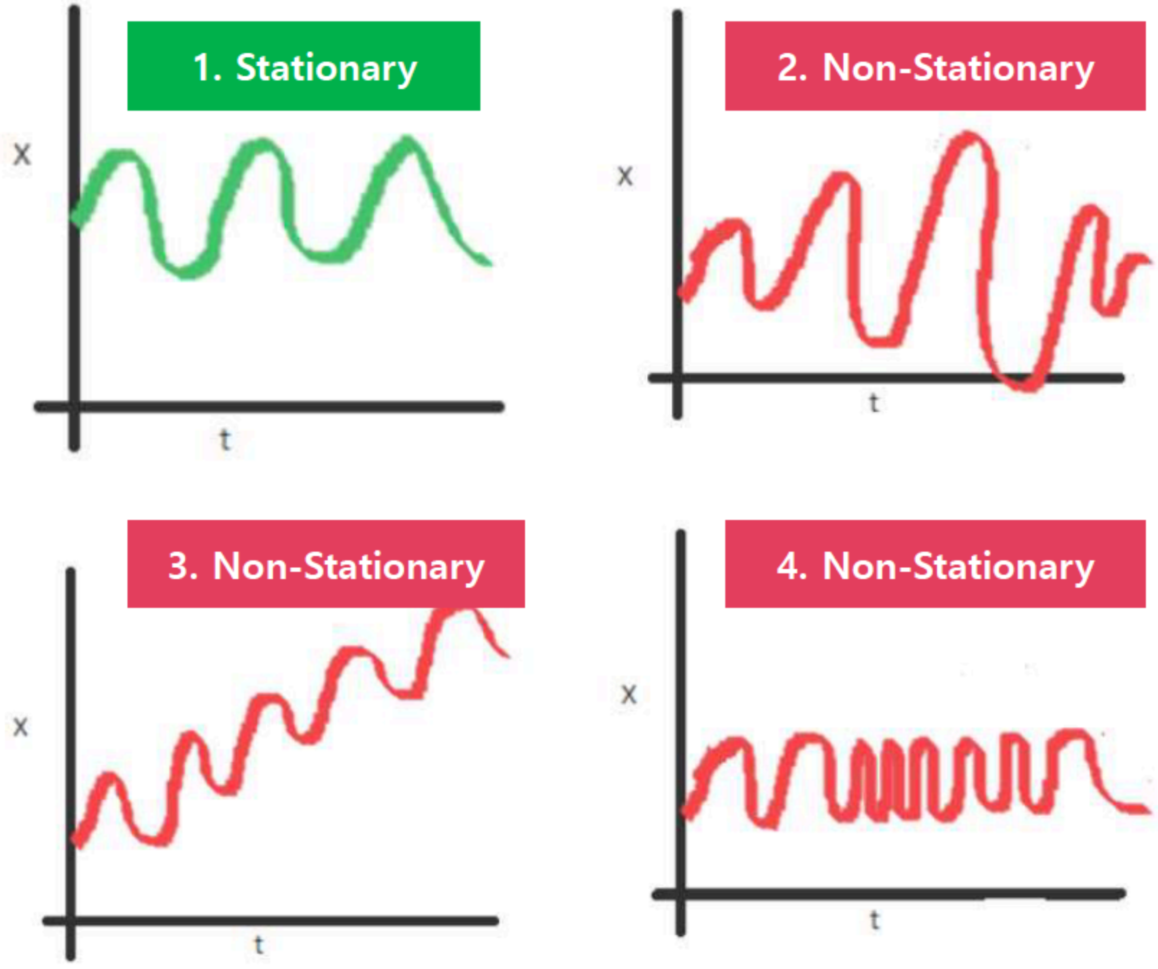
- **횡단면 분석(Cross-sectional analysis):** 주어진 시점에서 데이터들간의 관계를 이용한 회귀분석 및 예측.
- **시계열 분석(Time series analysis):** 순서가 있는 데이터의 분석 및 예측.
- 단순히 날짜와 시간이 있는 데이터라 해서 시계열 데이터가 아님!





정상 시계열 (Stationary time series)

stationary vs. non-stationary



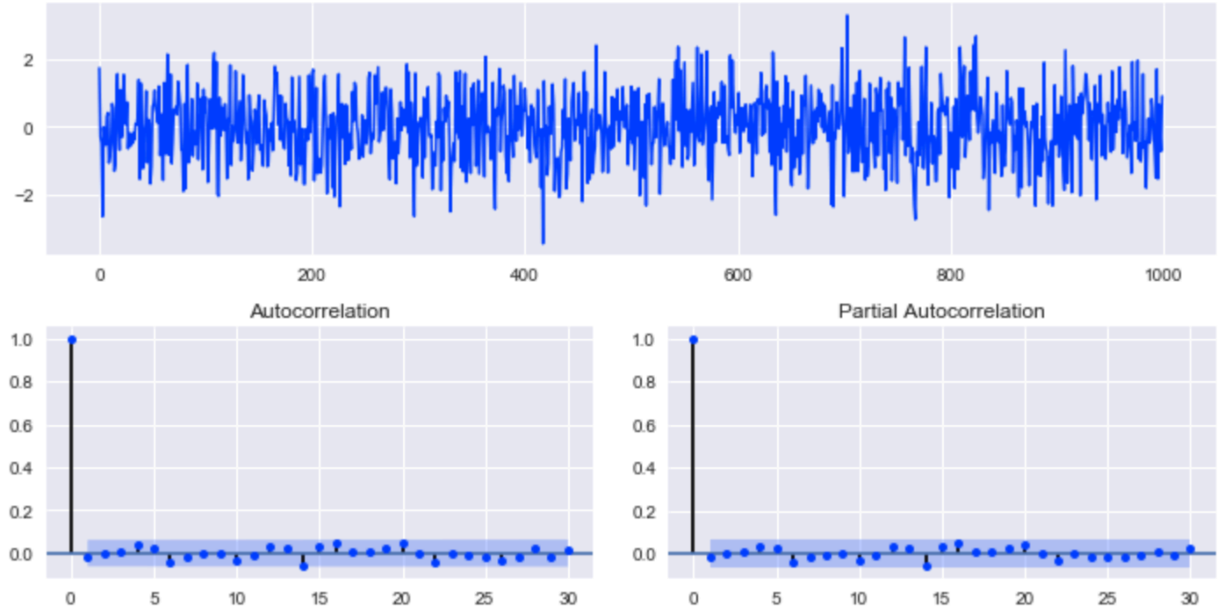
- **Strictly stationary** : $\{x_{t1},x_{t2},\ldots,x_{tk}\}$ is identical to $\{x_{t1+h},x_{t2+h},\ldots,x_{tk+h}\}$

- **Weakly stationary :**

  $E(x_t)$ = const, $Var(x_{t+h})=Var(x_t)=$const,

  $Cov(x_{t+h},\ x_{s+h})=Cov(x_t,\ x_s)$

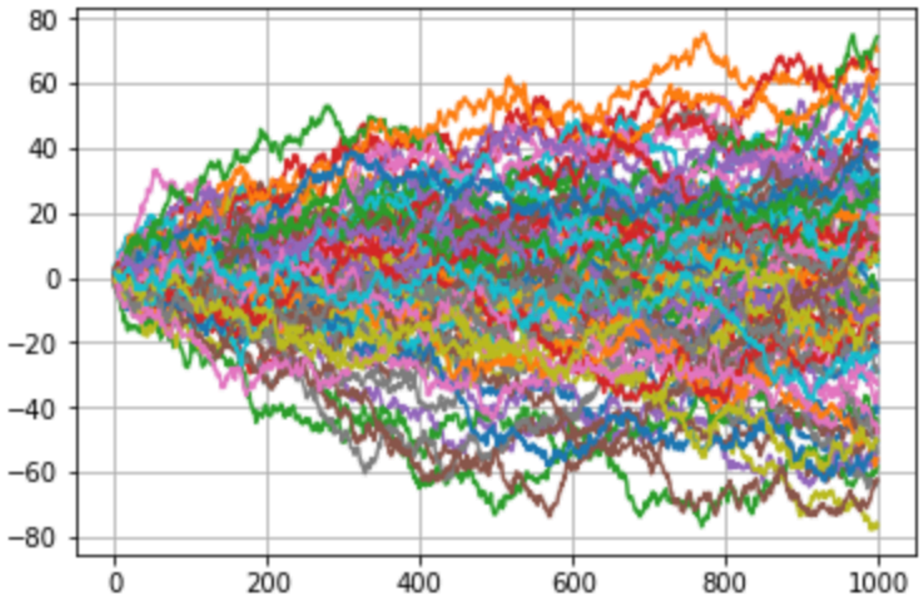
  (EX) Gaussian White Noise $w_t \sim N(0,\ \sigma^2)$



랜덤 워크(Random work)
대표적인 nonstationary 확률 과정

- **랜덤 워크(Random work):** White noise의 누적 확률 과정이다.
  - $X_t = \varepsilon_1 + \varepsilon_1 + \cdots + \varepsilon_t$
    $= X_{t-1} + \varepsilon_t$

  평균은 0이지만 분산은 시간이 지남에 따라 증가함.
  - $E[X_t] = 0,\ Var[X_t] = t\sigma^2$
  - 대표적인 nonstationary process

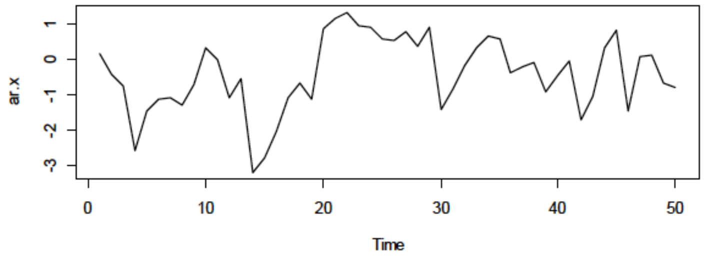
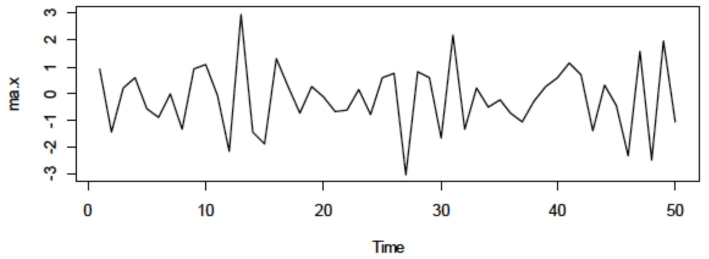
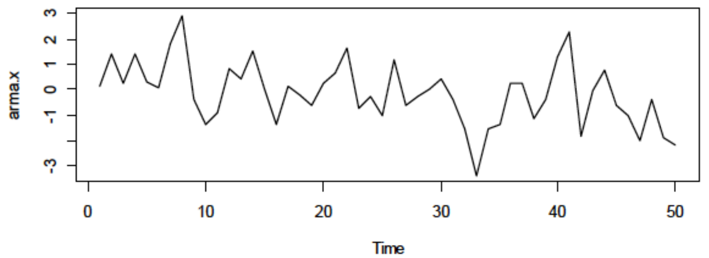


- **랜덤 워크의 차분:** 랜덤 워크를 차분하면 stationary하게 변환된다.
  - $\nabla X_t = X_t - X_{t-1} = \varepsilon_t \sim N(0, \sigma^2)$

- **시계열 분석의 workflow:**
  1. nonstationary -> stationary process를 추출 (차분, 추세 제거)
  2. 정규성 검정. 정규 분포가 아니라면 적절히 변환
  3. Stationary process에 대한 확률 모형 추정 (AR, MA, ARMA, ARIMA 등)
  4. 잔차에 대한 정규성 검정 (Ljung-Box Q test, QQ-plot, ACF) 및 예측력 검정(AIC, BIC)

일반 선형확률과정 모형(General linear process model)

대표적인 stationary 확률 과정

	데이터 구조	표현식	예시 그래프
Auto -regressive (AR)	▪ 예측에 활용하고 있는 데이터의 과거 데이터에 따라 값이 결정되는 데이터 프로세스 ▪ 어제의 결과가 오늘에 영향을 미치고, 오늘의 결과가 내일에 영향을 미치는 방식	$X_t = c + \sum_{i=1}^{p} \varphi_i X_{t-i} + \varepsilon_t .$	
Moving Average (MA)	▪ 과거에 예측했던 값에서 발생한 오차가 현재에 영향을 미치는 데이터 프로세스 ▪ 어제의 오차가 오늘에 영향을 미치고, 오늘의 오차가 내일에 영향을 미치는 방식	$X_t = \mu + \varepsilon_t + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i}$	
AR(I)MA	▪ 위의 두 데이터 프로세스가 결합된 경우 ▪ 어제의 결과와 어제의 오차가 오늘에 영향을 미치고, 오늘의 결과와 오늘의 오차가 내일에 영향을 미치는 방식 ▪ 많은 시계열 데이터가 AR(I)MA 프로세스를 따름	$X_t = c + \varepsilon_t + \sum_{i=1}^{p} \varphi_i X_{t-i} + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i}$	

자기회귀 모형(Autoregressive model)

AR(p)의 형태로 mean reversion을 설명한다.

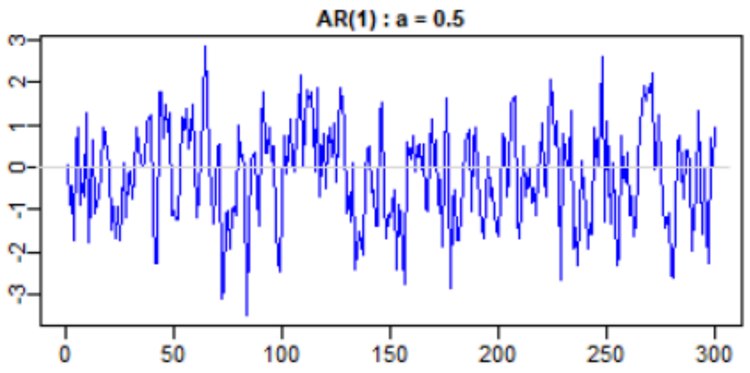
$$AR(p): \quad X_t = a_1 X_{t-1} + a_2 X_{t-2} + a_3 X_{t-3} + \cdots + \varepsilon_t$$

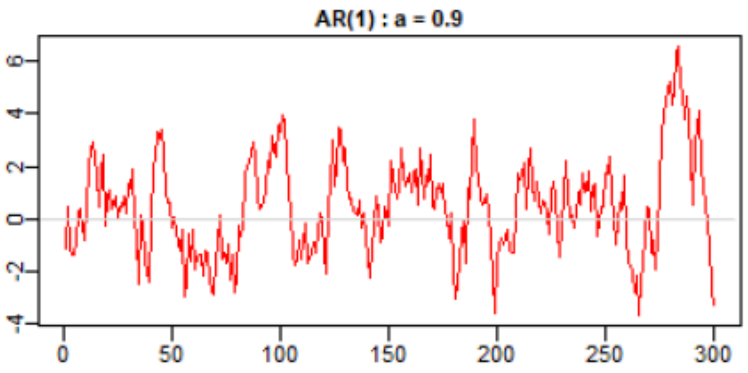
$$AR(1): \quad X_t = aX_{t-1} + \varepsilon_t \Rightarrow \mathrm{E}[X_t] = \mathrm{aE}[X_{t-1}] = \cdots = \mathrm{a}^{t-1}\,\varepsilon_1 = 0,\ \text{if } -1<a<1 \text{ (stationary condition)}$$

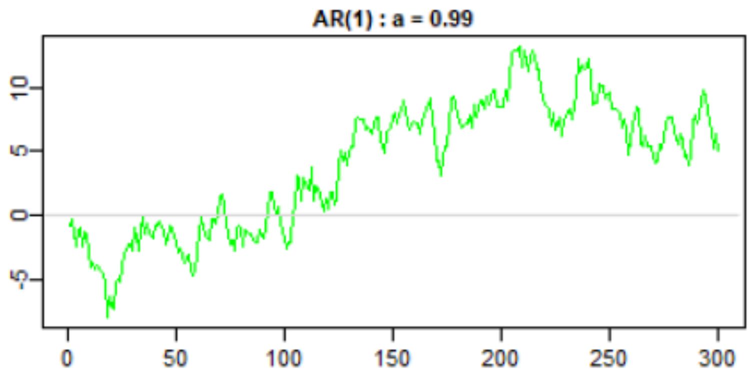
$$\mathrm{Var}[X_t] = \mathrm{E}[X_t{}^2] = \mathrm{E}[\mathrm{a}^2 \mathrm{X}_{t-1}{}^2 + 2\mathrm{aX}_{t-1}\varepsilon_t + \varepsilon_t{}^2] = \mathrm{a}^2\mathrm{Var}[X_{t-1}] + \sigma_e^2 = \sigma_e^2 + \mathrm{a}^2\sigma_e^2 + \mathrm{a}^4\sigma_e^2 + \cdots$$

$$= \frac{\sigma_e^2}{1-a^2},\ \text{if } -1<a<1 \text{ (stationary condition)}$$

- AR 모형의 특징
  1. a의 절대값이 1보다 작으면 weakly stationary 시계열이다
  2. a의 절대값이 1이상이면 random work 시계열이다
  3. a가 -1에 가까울수록 평균회귀 성향이 강하고, 1에 가까울수록 random work에 가깝다.







자기회귀 모형(Autoregressive model)

ACF, PACF특성을 분석하여 적합한 모형을 찾을 수 있다.

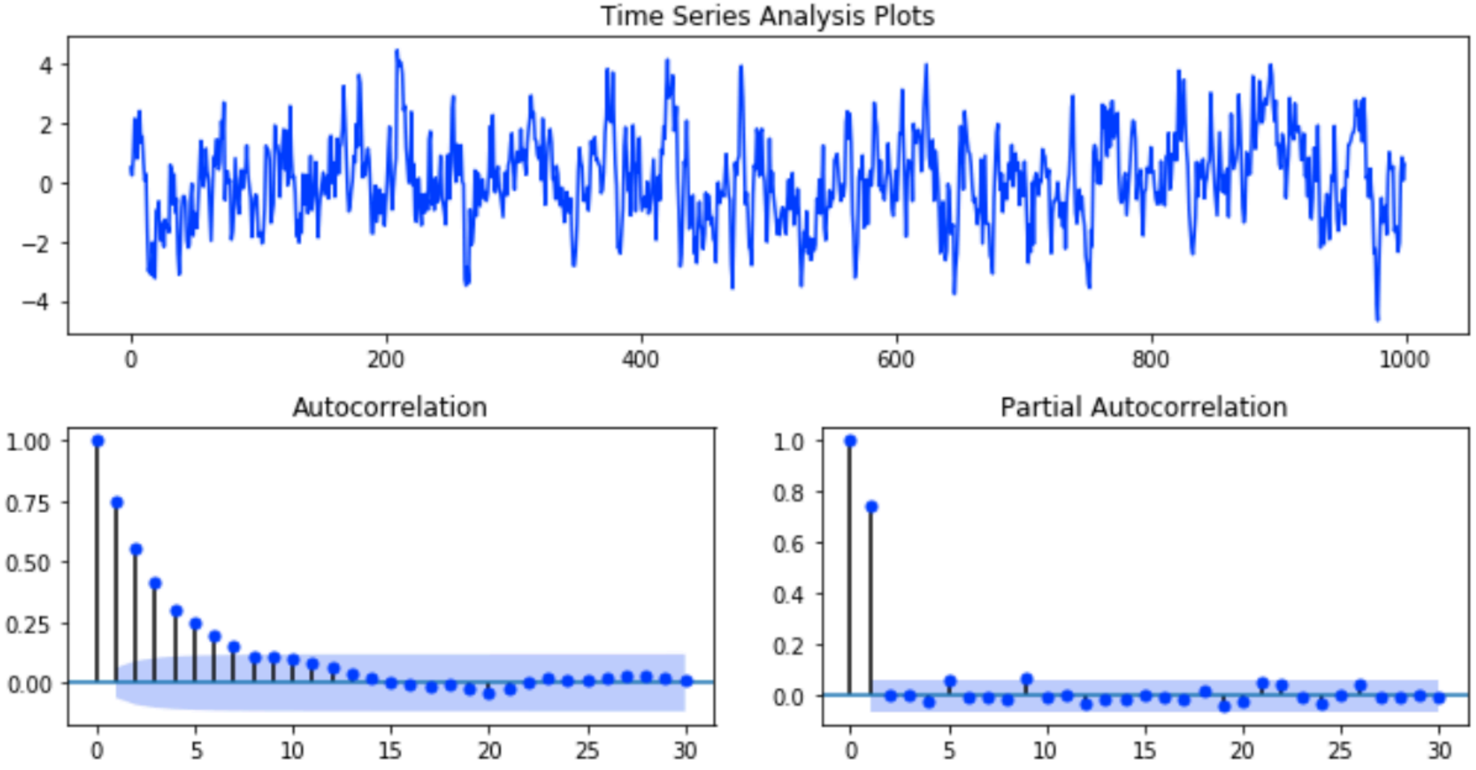
- ***Autocorrelation function (ACF)*** *: t와 t-h 데이터 사이의 자기공분산*

$$\gamma_h = E[X_t X_{t-h}] = E[aX_{t-1}X_{t-h} + \varepsilon_t X_{t-h}] = a\gamma_{h-1} = \cdots = a^h \frac{\sigma_e^2}{1-a^2} \text{ (exponential decay if stationary)}$$

- ***Partial Autocorrelation function (PACF)*** *: t와 t-h 외 중간 데이터들(t-1 … t-h+1)의 영향을 제외한 자기상관계수*

$$\rho_{11} = Corr[X_t, X_{t-1}] = a$$

$$\rho_{22} = Corr[X_t, X_{t-2} | X_{t-1}] = Corr[X_t - \alpha X_{t-1}, X_{t-2} - \beta X_{t-1}] = Corr[\varepsilon_t + function\ of\ \varepsilon_{t-1}] = 0$$



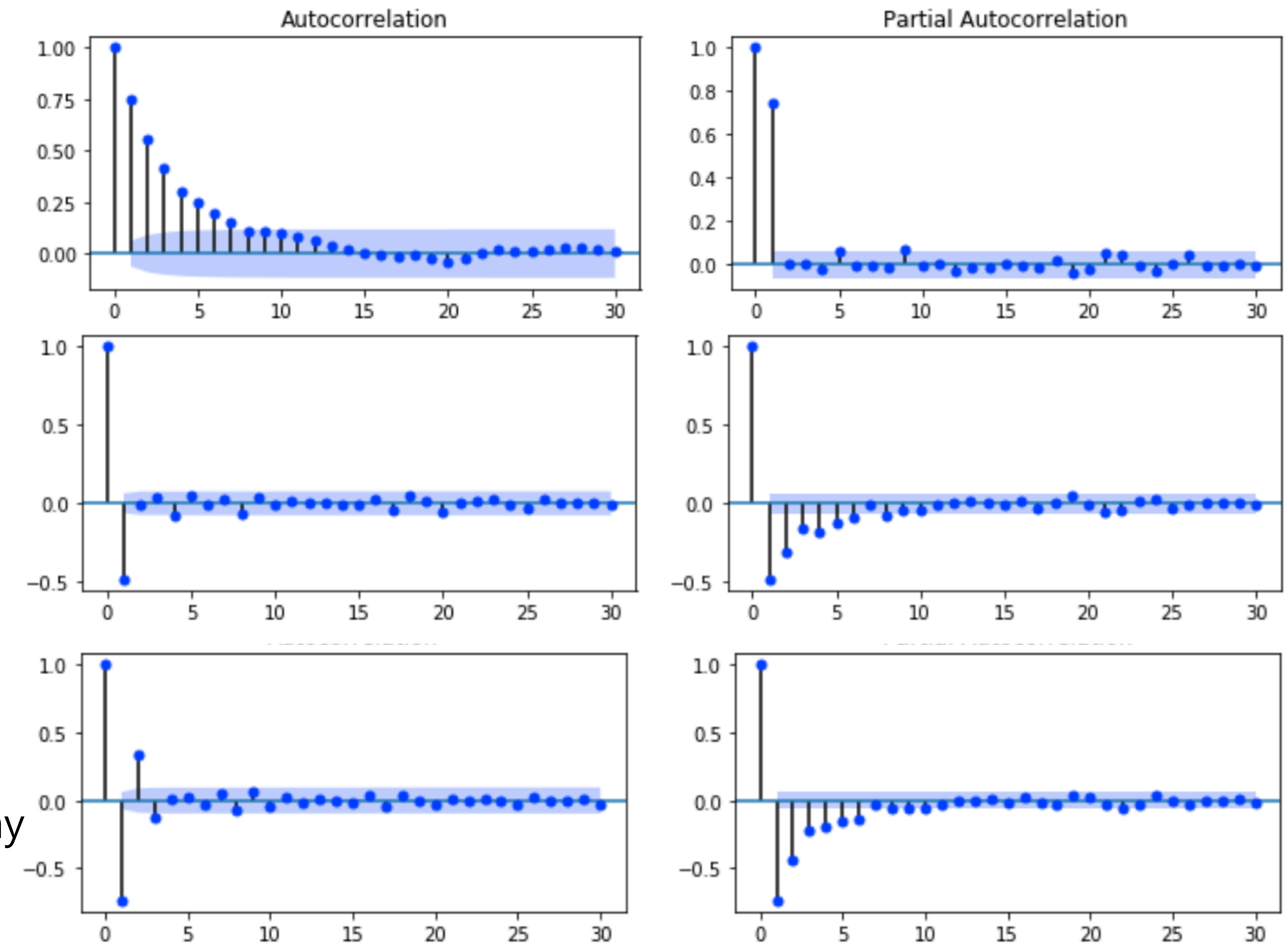
선형확률과정 모형의 Correlogram 비교

Correlogram 분석을 통해 적합한 AR, MA, ARMA모형을 찾을 수 있다.

AR 모형
exp. decay & break

MA 모형
Break & exp. decay

AR 모형
exp. decay & exp. decay



Box-Jenkins 방법과 모델의 결정

Stationary 시계열에서 최적의 모델을 결정하는 방법.

Box-Jenkins 방법

1. 데이터가 stationary 한지 검정, 또는 stationary 하게 transform된 데이터를 준비
2. ACF, PACF를 통해 가능한 모형의 후보군 선택
3. OLS 또는 Yull-Walker 방정식, 또는 conditional maximization (like MLE)기법을 통해 모수 추정
4. Diagnostic check 방법으로 자격 미달의 후보들을 차례대로 탈락 -> 최적의 모델 결정
5. Prediction

Diagnostic check 방법: 잔차항이 white noise인지 검정한다.

(1) Ljung-Box-Pierce Q test : $Q_* = T(T+2)\sum_{h=1}^{k}\frac{\hat{\gamma}_h}{T-h} \sim \chi^2(k)$ for ARMA(p, q) process

(여기서는 원 데이터의 자기상관계수를 이용함)

(2) 모형을 추정하고 남은 잔차에 대한 Q-test도 할 수 있음. 이때는 $\chi^2(k-p-q-1)$분포를 따름

(3) 그 외 CI measure, Jarque-Bera test, QQ plot, Sign test, Rank test 등이 있음

Prediction power의 검정: parameter 추가에 따른 패널티를 줘서 예측력을 평가한다.

(1) Akaike Information Criterion (AIC): $-2\log(L) + \frac{2(p+q+1)n}{n-p-q-2}$

(2) Bayesian Information Criterion (BIC): $-2\log(L) + 2(p+q+1)\log(n)$

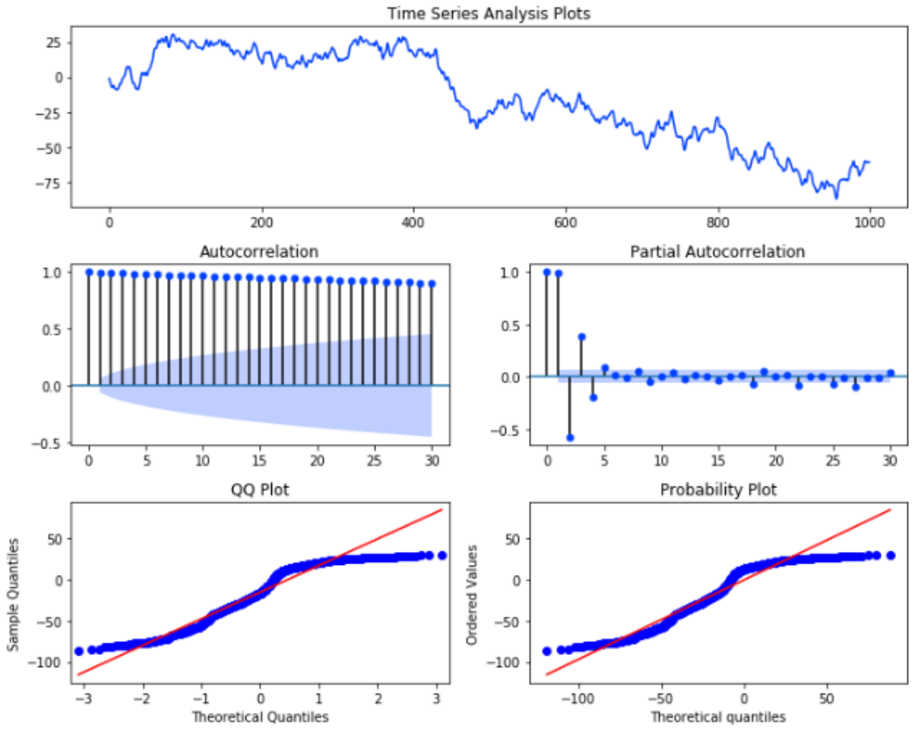
(L은 잔차가 얼마나 정규분포에 가까운지를 알려주는 likelihood, 따라서 AIC와 BIC는 작을수록 우수한 모델임을 뜻함)

Nonstationary process의 분석 – ARIMA 모형

nonstationary 시계열을 차분하여 ARMA 모형을 구축하면 ARIMA모형이 된다.

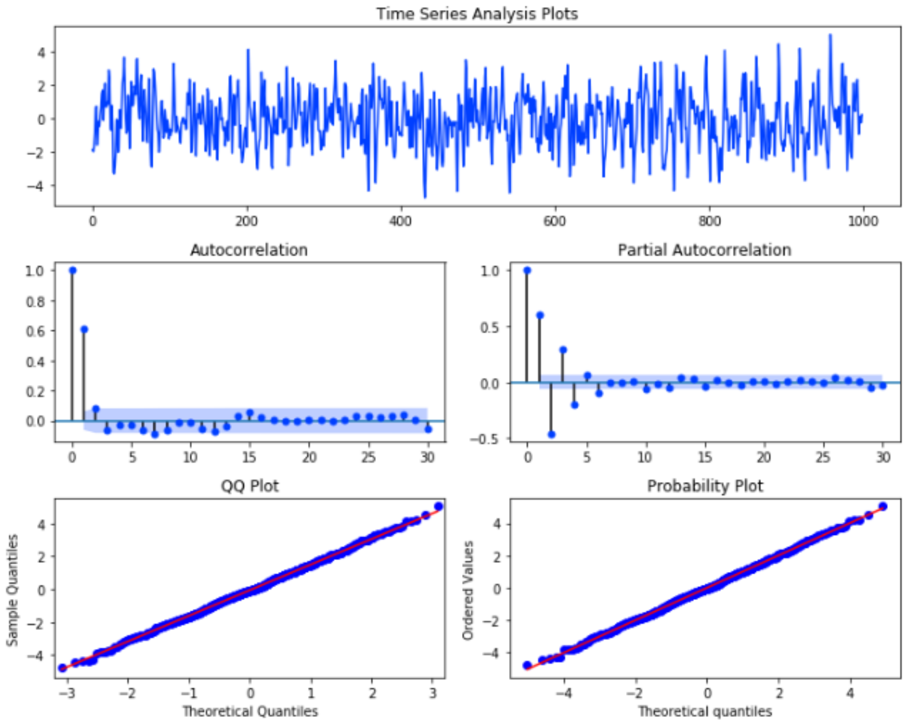
ARIMA(Autoregressive Integrated Moving Average) 모형:

- 시계열 X_t 를 d번차분한 $\nabla^d X_t = X_t - X_{t-1}$ 이 ARMA(p, q)를 따르면 X_t는 ARIMA(p,d,q) 모형을 따른다.
- 원 시계열의 추세 성분만 제거되었을 뿐, 원 시계열의 autocorrelation이나 volatility는 그대로 유지된다.
- 즉 구간에 따라 수준 (level)은 다를지라도, 그 행태가 수준에 관계없이 유사성을 가지는 등질비정상 시계열에 적합함.
  (비등질시계열의 경우는 단순차분보다는 log를 취하거나 Box-Cox transformation 등을 하여 정규분포에 가깝게 만드는게 좋다)



Difference

Integration



ARIMA(2,1,1)

ARMA(2,1)

단위근 검정 (Unit root test)

DF, ADF test를 통해 시계열의 stationary 여부를 검정할 수 있다.

Dickey-Fuller 단위근 검정 (DF test)

- $Y_t = a_1 Y_{t-1} + a_2 Y_{t-2} + \cdots + a_p Y_{t-p} + \varepsilon_t$
- 위 식을 변형하면, $Y_t = \rho Y_{t-1} + b_1 \nabla Y_{t-1} + b_2 \nabla Y_{t-2} + \cdots + b_p \nabla Y_{t-p} + \varepsilon_t$ $(\rho = \sum a_k)$
- unit root를 가진다는 것은 $\sum a_k = 1$ 임. $\Rightarrow$ $\begin{cases} H_0: & \rho = 1 \\ H_1: & \rho < 1 \end{cases}$

  $\hat{\rho}$는 OLS로 구함 $\Rightarrow \hat{\rho} = \frac{\sum_{t=1}^{T} Y_{t-1} Y_t}{\sum_{t=1}^{T} Y_{t-1}^2}$

- 귀무가설 하에서, $\mathrm{T}(\hat{\rho} - 1)$은 asymptotically 아래 분포함수(시뮬레이션을 통해 얻음)를 따르고, 그에 따라 t-검정하면 됨.

  $\mathrm{T}(\hat{\rho} - 1) \Rightarrow \frac{\int_0^1 W(r) dW(r)}{\int_0^1 W(r)^2 dr}$, $W(r)$은 the Wiener process (or Brownian process)

Augmented Dickey-Fuller 단위근 검정 (ADF test)

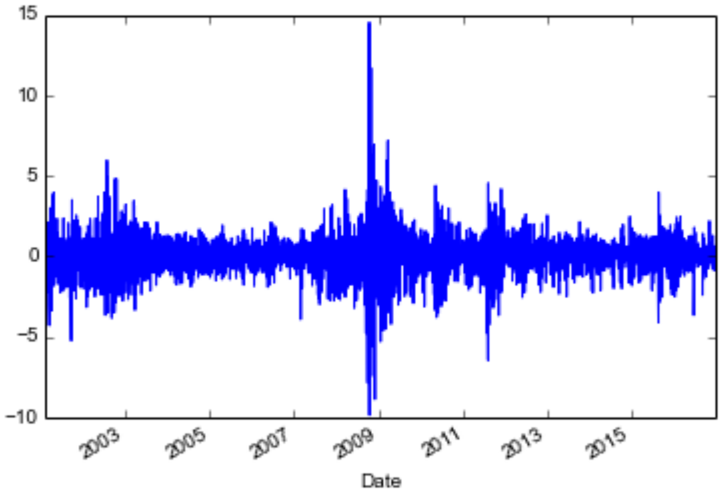
- DF test는 Y_t에 대한 regression인 반면 ADF test는 ∇Y_t에 대하여 regression
- 절편항과 추세성분도 포함됨
- $\nabla Y_t = \rho Y_{t-1} + \alpha + \beta t + b_1 \nabla Y_{t-1} + b_2 \nabla Y_{t-2} + \cdots + b_p \nabla Y_{t-p} + \varepsilon_t$
- DF test와 유사하게 t-statistics를 구축하여 검정.
- $\rho < 0$ 은 mean reversion을 뜻한다.

Conditional Heteroskedastic Models (Time-Varying Volatility)

ARCH, GARCH 모형을 통해 시간에 따른 변동성의 변화를 모델링 할 수 있다.

AutoRegressive Conditional Heteroskedasticity (ARCH) by Engle (1982), Nobel Prize (2003)

금융 시계열의 time-varying volatility를 모델링 하기 위하여 만들어진 모형
$X_t = aX_{t-1} + \varepsilon_t$에서 ε_t가 White Noise (w_t)가 아닐 때 사용한다.
ε_t에 대한 correlogram은 white noise 처럼 보일 수 있지만 통상 ε_t^2에 autocorrelation이 발생한다.



- *ARCH(1)*

  $\varepsilon_t = \sigma_t w_t$ where σ_t is called volatility given by $\sigma_t^2 = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2$

  $Var(\varepsilon_t) = \alpha_0 + \alpha_1 Var(\varepsilon_{t-1})$

- *ARCH(q)*

  $\varepsilon_t = \sigma_t w_t$ where σ_t is called volatility given by $\sigma_t^2 = \alpha_0 + \sum_{i=1}^{q} \alpha_i \, \varepsilon_{t-i}^2$

Generalized AutoRegressive Conditional Heteroskedasticity (GARCH)

- *GARCH(p, q)*

  $\varepsilon_t = \sigma_t w_t$ where σ_t is called volatility given by $\sigma_t^2 = \alpha_0 + \sum_{i=1}^{q} \alpha_i \, \varepsilon_{t-i}^2 + \sum_{j=1}^{p} \alpha_i \, \sigma_{t-j}^2$

- 통상 GARCH(1,1)이면 금융 시계열 모델링에 충분

Time Series Model Develop for Financial Data

- **ARIMA + GARCH model building**

  1. Asset price => Return으로 바꿈.
  2. Stationary Test (ADF)
  3. ARMA or ARIMA fitting
  4. Best model selection by Box-Jenkins method
  5. GARCH fitting to residuals
  6. Normality test & ADF test for residuals of (ARIMA + GARCH) model.

- **Further issues**

  1. Seasonal ARIMA(SARIMA), ARX, ARFIMA, VAR, VECM등의 여러 모델
  2. Spectral analysis
  3. State-space model (Kalman filtering, Hidden Markov Model, etc.)
  4. Combining Machine Learning technique
  5. Model-free data based approach
     (Recurrent Neural Network (RNN), Long-Short Term Memory (LSTM), etc.)