Data Science in Financial Markets CSE4009

Dr. Hirdesh Kumar Pharasi
Associate Professor
BML Munjal University, Haryana, India

1

Overview of Financial Markets

Stock Markets:

- Platforms for buying and selling shares of public companies.
- Examples include the New York Stock Exchange (NYSE) and NASDAQ.

Bond Markets:

- Markets for trading debt securities.
- Governments, municipalities, and corporations issue bonds to raise capital. Examples: US Treasury bonds traded in the US bond market

Commodity Markets:

- Platforms for trading physical goods like gold, oil, agricultural products, etc.
- Examples include the Chicago Mercantile Exchange (CME) and London Metal Exchange (LME).

Derivatives Markets:

- Markets for trading financial contracts like futures and options.
- Derivatives derive their value from underlying assets like stocks, bonds, or commodities.

2

1. Futures Contracts

- **Definition:** Standardized agreements to buy or sell an asset at a predetermined price on a set future date, traded on exchanges (e.g., NSE, BSE).
- **Example**: Oil futures are standardized contracts traded on NYMEX where buyers and sellers agree to trade a specific number of barrels of crude oil at a set price on a future date. Regardless of the market price then, **both parties are obliged to transact at this price**.
- **Use case**: Hedging against price fluctuations (as in the Oil example) or speculating on future price movements (such as buying Nifty50 index futures).

2. Options Contracts

- **Definition**: Give the holder the right, but not the obligation, to buy (call option) or sell (put option) an asset at a specified price within a certain time frame, in exchange for a premium.
- **Example**: An investor buys a call option to purchase Reliance Industries shares at ₹2,800 anytime in the next month. If the share price rises above ₹2,800, the investor can buy at the lower price; if not, the investor lets the option expire and **only loses the premium paid**.
- **Use case**: Limiting risk to the premium (for buyers), generating income through option writing, hedging positions, or speculating on market moves.

3

3. Forwards Contracts

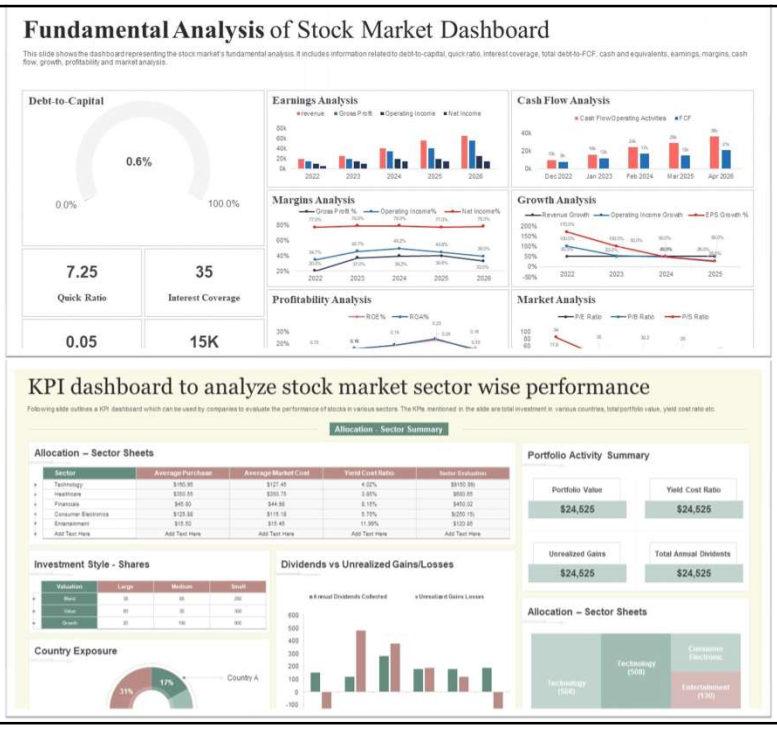
- **Definition**: Customized, privately negotiated agreements to buy or sell an asset at a future date for a pre-agreed price, traded over-the-counter (OTC).
- **Example**: An exporter enters into a forward contract with a **bank** to sell US$1 million at ₹83 per dollar in six months, thus hedging against currency risk if the rupee value fluctuates later.
- **Use case**: Tailoring agreements to specific needs (amount, date, underlying asset), mainly for hedging. Less standardized and less liquid than futures.

4. Swaps

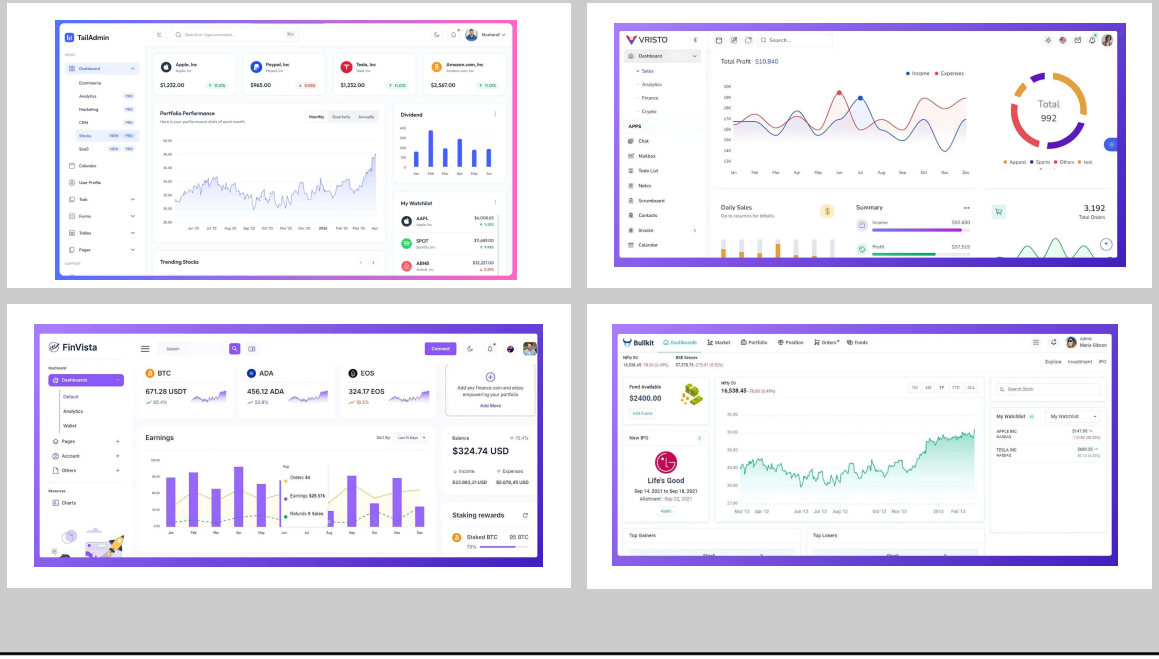
- **Definition**: Agreements between two parties to exchange cash flows or other financial instruments, commonly to manage interest rate or currency risks.
- **Example**: Two companies, one with a fixed-interest loan and another with a floating-interest loan, swap their interest payment streams. This lets both achieve more favorable borrowing conditions.
- **Use case**: Managing interest rate exposure (interest rate swaps), accessing foreign currencies (currency swaps), or altering cash flow structures.

4

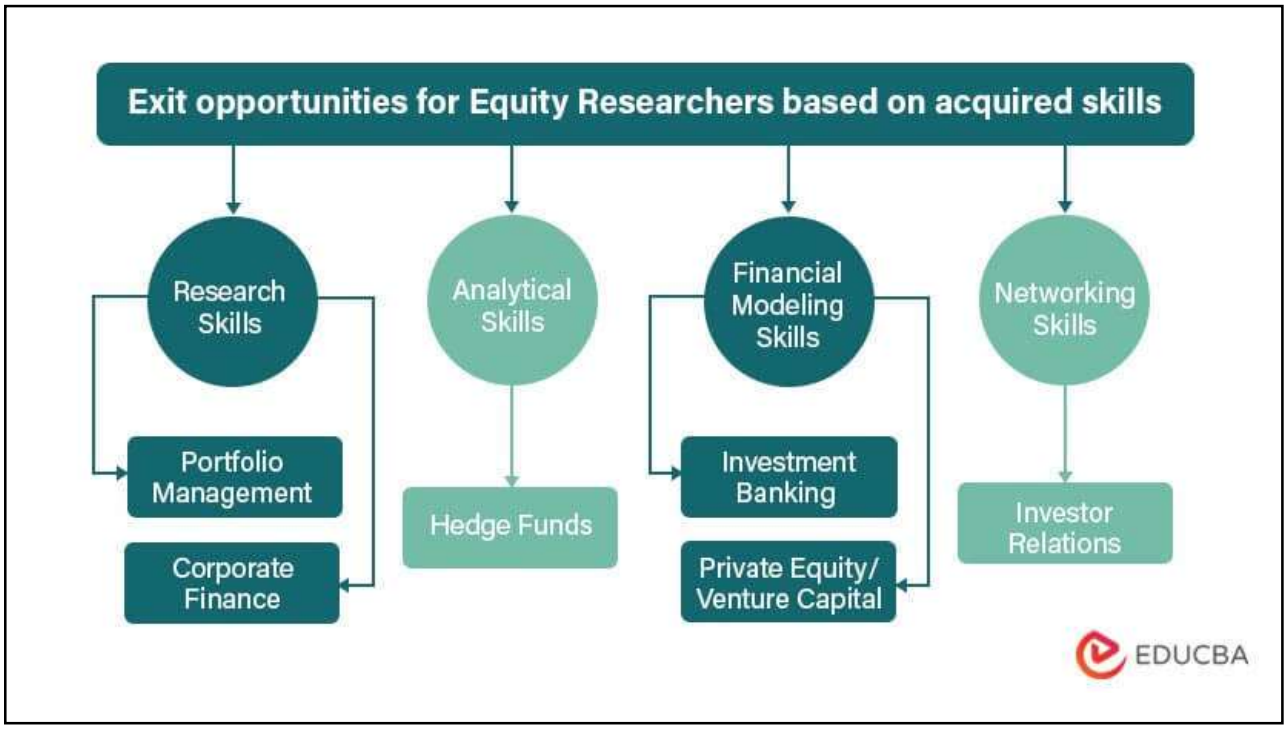
Stock Market Dashboard



5



6



7

Job Opportunities After Data Science in Financial Markets

- **Core Roles**
  - **Financial Data Scientist**: Predictive modeling, algorithmic trading, big data analytics
  - **Quantitative Analyst (Quant)**: Pricing, risk models, strategy development
  - **Data Analyst (Finance):** Data extraction, trend analysis, business/credit insights
  - **Risk Analyst / Manager**: Market/credit risk modeling, stress tests
  - **Machine Learning Engineer (Finance):** ML/DL for fraud, trading, and credit scoring
  - **Business Intelligence Analyst**: Dashboards, reporting, BI tool expertise
  - **Data Engineer / Architect**: Large-scale financial data pipelines
  - **Investment Banking Analyst**: Analytics for IPOs, capital markets

8

Industries, Specialized Roles & Outlook

- **Specialized/Emerging Roles**
  - Financial Research Analyst: Data-driven economic/investment research
  - Credit Analyst: Loan portfolio analytics, credit scoring
  - Portfolio Analytics Specialist: Portfolio optimization, stress testing
  - Fintech Product Analyst: Product improvement, analytics in fintechs
  - HFT Strategist: Microstructure, real-time trading algorithms
  - NLP Expert (Finance): Sentiment, event extraction from text, news, and filings
- **Industries/Employers**
  - Investment & retail banks, Asset managers, Hedge & quant funds
  - Fintech startups, Consulting, Insurers, Regulators & exchanges
- **Growth**
  - Strong demand for talent blending finance, data science, and analytics
  - Roles evolving fast with AI, big data, and quantitative modeling in finance

9

CFM Capital Fund Management

Client login

Jean-Philippe Bouchaud

Chairman

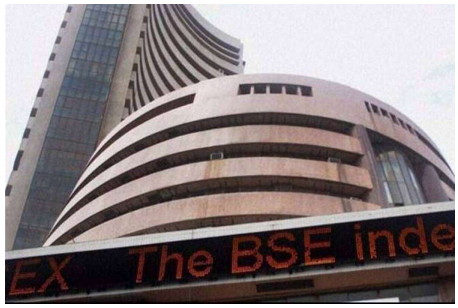
Jean-Philippe is Chairman of CFM. He founded 'Science and Finance' in 1994, the research arm of CFM with Jean-Pierre Aguilar, which merged with CFM in 2000. He supervises the research team alongside Marc Potters. Jean-Philippe maintains strong links with the academic world and is a professor at École Normale Supérieure (ENS). Prior to CFM, Jean-Philippe was a researcher at the Centre National de la Recherche Scientifique until 1992. Following this he spent a year at the Cavendish Laboratory in Cambridge before joining the Service de Physique de l'État Condensé at the Commissariat à l'Energie Atomique in Saclay, France. He holds a PhD in theoretical physics from the ENS in Paris.

CFM has approx. 350 employees worldwide and manages $15 billion as of Aug 2024.

10

Financial Markets

Financial market:
Collective behavior of interacting agents.
Agents interact to perform a collective task of finding the **best price for a financial asset**.



11

What Are Financial Markets?

- Definition: Platforms or systems where participants buy and sell financial assets (stocks, bonds, commodities, derivatives).
- Purpose: Facilitate raising capital, investment, liquidity, and price discovery.
- Examples: Stock exchanges (NYSE, Nasdaq), commodity markets, bond markets.

Types of Financial Markets

- **Primary Market**
  - Where new securities are issued for the first time.
  - Example: Initial Public Offerings (IPOs).
  - Funds go directly to issuers (companies/governments).
- **Secondary Market**
  - Where previously issued securities are traded.
  - Example: Stock exchanges, over-the-counter (OTC) markets.
  - Prices are set by supply and demand among investors.

12

Market Participants

- **Individual (Retail) Investors**: People managing their own investments.
- **Institutional Investors**: Entities like mutual funds, pension funds, insurance companies handling large sums.
- **Brokers & Dealers**: Facilitate buy/sell transactions; dealers may trade on their own account.
- **Market Makers**: Ensure liquidity by continuously quoting buy/sell prices.
- **Regulators**: Organizations (e.g., SEBI, SEC) that oversee fair and efficient market functioning

13

Notable Large-cap stock indices

Market (Country)	Stock Exchange	Benchmark Index	Index Constituents	Key Features
India	Bombay Stock Exchange (BSE)	Sensex (S&P BSE Sensex)	30 large-cap companies	Oldest Indian exchange; sector-representative; **market cap weighted** based on free-float shares
	National Stock Exchange (NSE)	Nifty 50 (NSE Nifty)	50 large-cap companies	Broader base than Sensex; 24 sectors; highly liquid; **free-float** market cap weighted
USA	New York Stock Exchange (NYSE)	S&P 500	500 large-cap companies	Widely followed US market benchmark representing large-cap US equities
USA	NASDAQ	NASDAQ Composite	3000+ companies (tech-heavy)	Heavy tech focus; includes many growth and tech stocks
UK	London Stock Exchange (LSE)	FTSE 100	100 largest companies	UK blue-chip companies; market cap weighted
Japan	Tokyo Stock Exchange (TSE)	Nikkei 225	225 large companies	Price-weighted index; reflects Tokyo market
Germany	Frankfurt Stock Exchange	DAX	40 blue-chip companies	Germany's leading index; market cap weighted
Canada	Toronto Stock Exchange (TSX)	S&P/TSX Composite	250 companies	Canada's principal stock index

14

Aspect	Market Cap Weighted	Free-Float Market Cap Weighted
Shares Considered	All outstanding shares	Only publicly tradable (free float) shares
Includes	Locked-in shares (promoters, govt.)	Excludes locked-in or restricted shares
Purpose	Measures company size	Measures company size adjusted for actual trading availability
Index Examples	Dow Jones Industrial Average (some cases)	S&P 500, Nifty 50
Impact on Weightings	Large promoter holdings can inflate weight	Reflects liquidity better and reduces influence of illiquid shares

15

Definition

Large-cap stocks are shares of companies with a *large market capitalization*, typically representing the most established and financially strong firms in the market.

Definition & Criteria:

Market capitalization (market cap) = Current share price × Total outstanding shares.

Large-cap stocks usually refer to companies with market caps above:

- ₹20,000 crore and above in India (top 100 companies by market cap).
- $10 billion and above globally

Feature	Explanation
Large Market Cap	Top companies by market cap, large financial footprint
Stability & Size	Well-established companies, lower risk, steady earnings
Dividend Paying	Often pay regular dividends
High Liquidity	Frequently traded, making entry/exit easy
Blue-chip Reputation	Trusted, dominant market players

16

Notable small-cap stock indices

Country	Index Name	Exchange / Provider	No. of Constituents	Market Cap Range / Coverage	Key Characteristics
UK	MSCI UK Small Cap Index	MSCI	~206	Covers ~14% of UK free float market cap; market cap approx. $166M to $9B	Free-float market cap weighted; diversified small-cap coverage; growth & liquidity focused
	FTSE Small Cap Index	London Stock Exchange (LSE)	Varies (subset)	Small-cap segment of UK equities	Market cap weighted; captures growth potential & dividends in small caps
USA	Russell 2000 Index	FTSE Russell	~2000	US small-cap stocks typically below $3B market cap	Widely followed US small-cap benchmark; free-float market cap weighted; diverse and liquid
Canada	S&P/TSX SmallCap Index	Toronto Stock Exchange (TSX)	Varies (~small-cap)	Smaller Canadian stocks relative to broader TSX	Market cap weighted; diversified; reflects emerging Canadian companies
Japan	S&P Japan SmallCap 250	S&P Dow Jones / Tokyo Exchange	250	Small-cap equities from Tokyo, Osaka, JASDAQ	Market cap weighted; liquidity focused; broad small-cap coverage
India	Nifty Smallcap 50	NSE	50	Top 50 companies from top 150 small caps by free-float market cap	Free-float market cap weighted; semi-annual rebalancing; focused on top small caps; high liquidity within small-cap segment.

17

What do we observe?

- Time Series of prices
- Transaction prices, best bid, best ask, spreads, orderbook, etc.

18

Market Crash

19

Time Series

20

Introduction to Time Series

Definition, Types, and Differences from Traditional Analysis

21

What is a Time Series?

- A sequence of data points recorded at regular time intervals
- Each observation is time-stamped
- Order of data matters (not independent)

Examples:

- Daily stock prices
- Monthly sales
- Hourly temperature readings

Types of Time Series

1. Univariate Time Series:
   - Single variable over time (e.g., stock price)
2. Multivariate Time Series:
   - Multiple variables recorded at each time step
   - Example: Price, Volume, Volatility of a stock

22

How is Time Series Analysis Different?

Aspect	Time Series Analysis	Traditional Data Analysis
Time Dependency	Observations are ordered and dependent	Observations assumed independent
Order Matters	YES – chronological order is crucial	NO – order often irrelevant
Goal	Forecast future, detect patterns/anomalies	Classification, clustering, regression
Tools Used	ACF, PACF, ARIMA, GARCH, LSTM	Linear regression, decision trees, etc.
Stationarity	Needs special treatment	Not applicable
Feature Engineering	Lags, rolling stats, differencing	Derived features, encoding, etc.

23

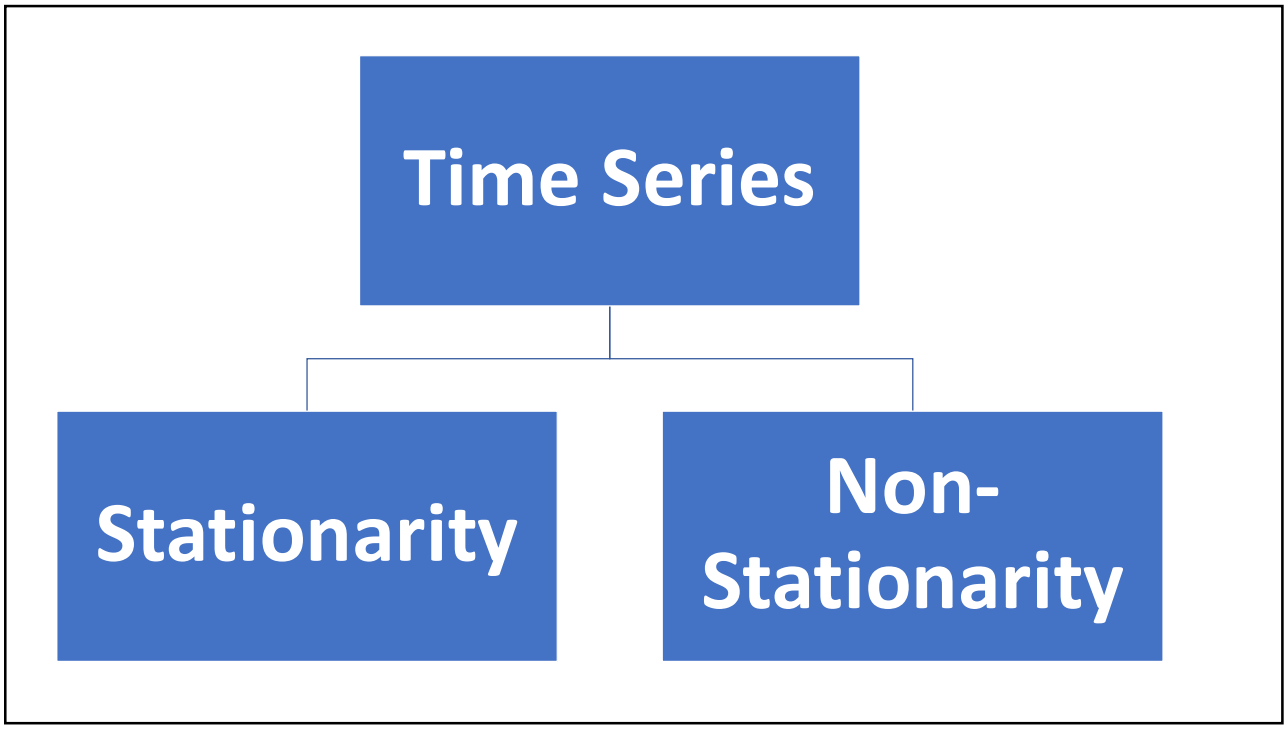
Key Concepts in Time Series

- Trend: Long-term upward/downward movement
- Seasonality: Repeating patterns over time
- Cyclic patterns: Irregular cycles (e.g., economic)
- Noise: Random variation not explained by models

Summary

- Time series are sequential and time-dependent
- Require unique modeling approaches
- Traditional models fail to capture time dynamics
- Foundation for forecasting and financial modeling

24



25

What is Stationarity?

A stationary time series has constant statistical properties over time

Important for model assumptions (ARIMA, GARCH, etc.)

Types

Strict Stationarity and Weak Stationarity

Strict Stationarity

- Joint distribution does not change over time
- All moments (mean, variance, skewness, Kurtosis.) are constant
- Mathematically: $P(X_{t1}, ..., X_{tk}) = P(X_{t1+h}, ..., X_{tk+h})$
- Example: White noise series with fixed distribution

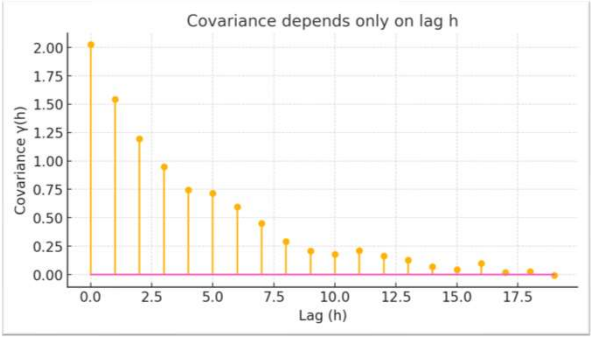
26

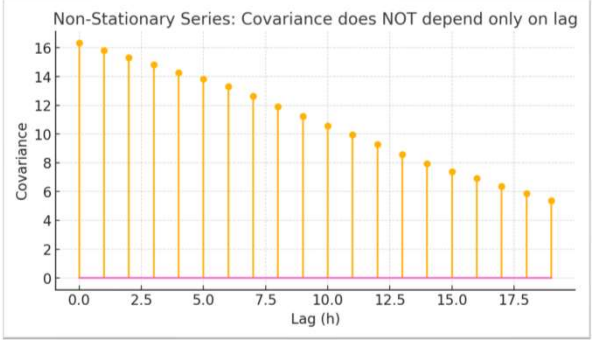
Weak Stationarity

- Only first two moments are time-invariant:
  - Constant mean $E[X_t]=\mu$ *(constant mean)*
  - Constant variance $Var(X_t)=\sigma^2$ *(constant variance)*
  - Covariance depends only on lag $Cov(X_t,X_{t+h})=\gamma(h)$ *(depends only on lag h)*
- Common in financial time series (e.g., stock returns)

Feature	Strict Stationarity	Weak Stationarity
Definition	Joint distribution doesn't change	Mean, variance, and covariance don't change
All moments constant?	All moments (1st, 2nd, …)	Only 1st and 2nd moments
Practical testing	Difficult (needs entire distribution)	Commonly used (via ADF test, etc.)
Required for models	Too strong, rarely used	Assumed in most models like ARIMA
Real-world examples	Rare (idealized cases only)	Common (e.g., stock returns)

27





- **Stationary series (previous plot) →** Covariance $\gamma(h)$ only depends on lag h and **decays consistently**.
- **Non-stationary series (random walk, above plot) →** Covariance does not just depend on h; it stays **large for long lags** and is affected by the actual time position t.

28

What Is Stationarity?

A stationary time series is a time series where there are no changes in the underlying system.

- Constant mean (no trend)
- Constant variance (no heteroscedacticity)
- Constant autocorrelation structure
- No periodic component (no seasonality)

29

Stationary Time Series

- (Weakly) stationary
  - The covariance is independent of t for each h

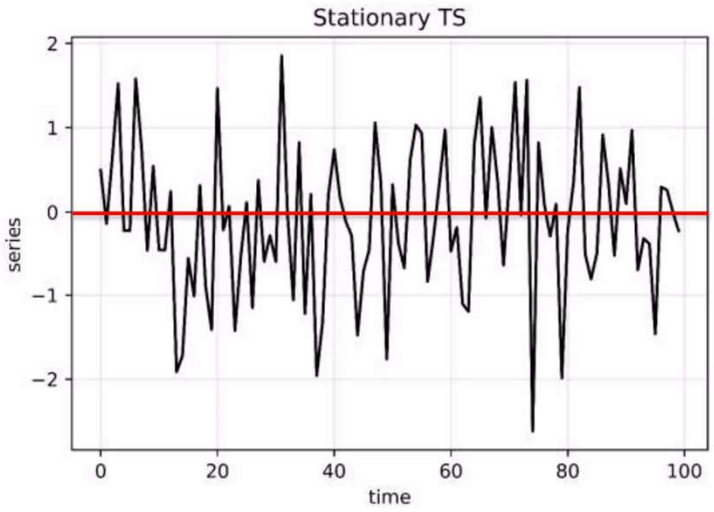
$$\gamma_X(X_t, X_{t-h}) \equiv E\left[(X_t - \mu)(X_{t-h} - \mu)\right]$$

  - The mean is independent of t

$$E(X_t) = \mu$$

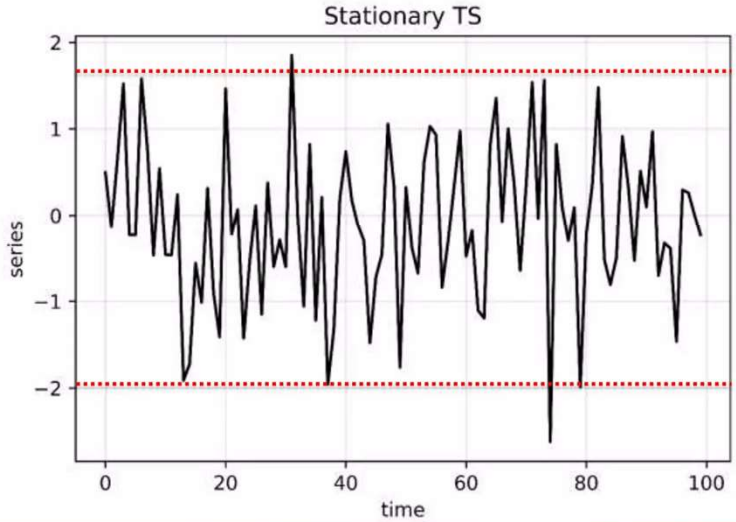
30

Assumption 1: Constant Mean



31

Assumption 2: Constant Variance



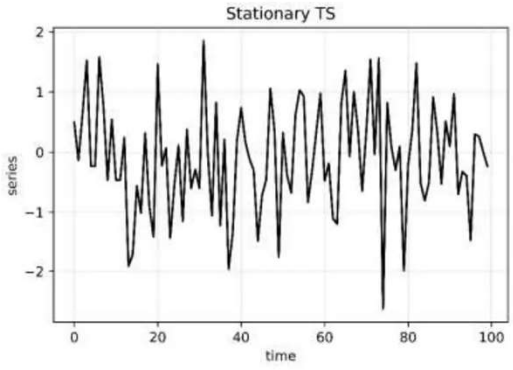
32

Autocorrelation

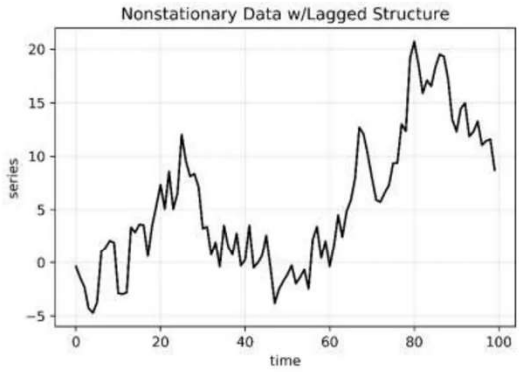
Autocorrelation is a key concept in time-series analysis.

- Autocorrelation is the correlation between a measurement at two different times.
- The time interval between values is called the lag.
- For example, stock prices may be correlated from one day to the next with a lag value of 1.
- Autocorrelation often results in a pattern, whereas a time series without autocorrelation will exhibit randomness.

33

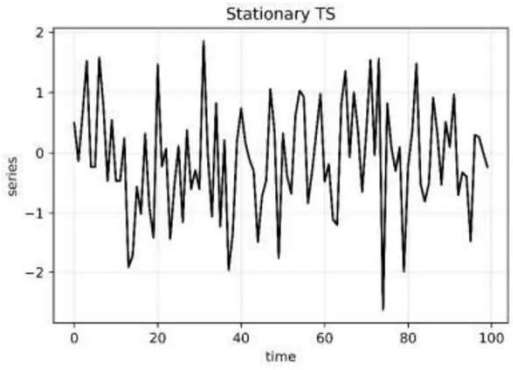


No autocorrelation

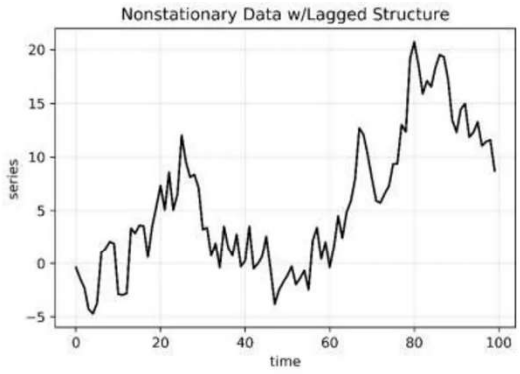


Has autocorrelation

34



No autocorrelation



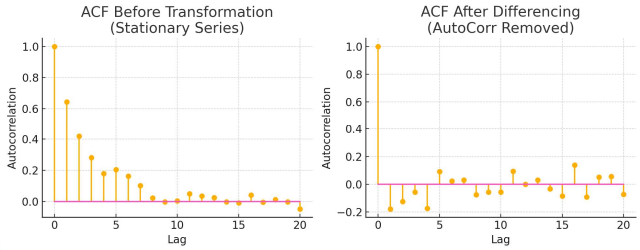
Has autocorrelation

35

Assumption 3: Constant Autocorrelation Structure

A stationary time series has constant autocorrelation structure throughout the entire series.

- If the autocorrelation remains constant throughout the series, a simple transformation can be used to remove the autocorrelation.
- This will be useful for several future models.



36

Why Is Stationarity Important?

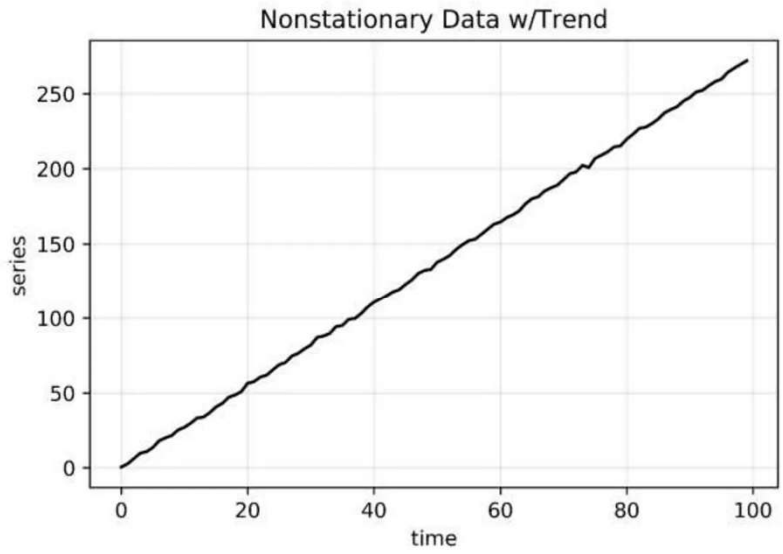
Stationarity is a fundamental assumption in many time-series forecasting models:

- Without it many basic time-series models would break down.
- Transformations can be applied to convert a nonstationary time series to a stationary one before modeling.
- While there are more advanced time-series models that can handle nonstationary data, that is beyond the scope of this lesson.

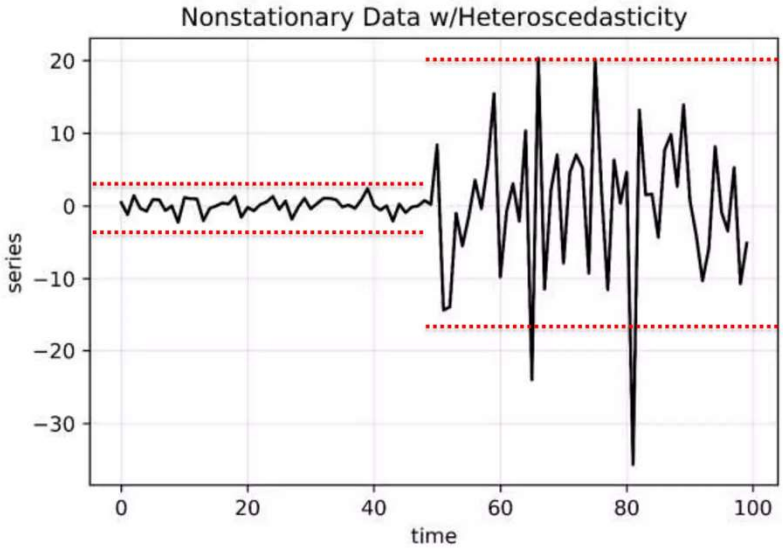
37

Nonstationary Examples

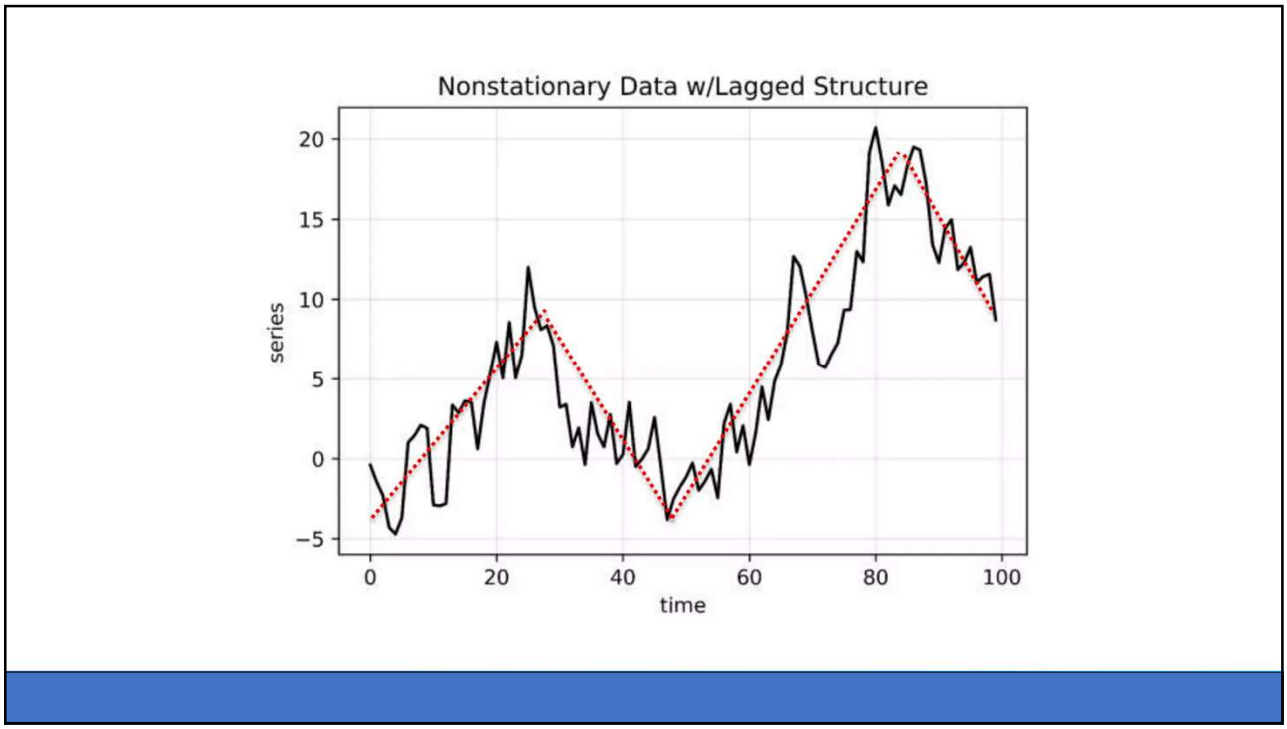
38


Nonstationary Data w/Trend
series
time
0
50
100
150
200
250
0
20
40
60
80
100

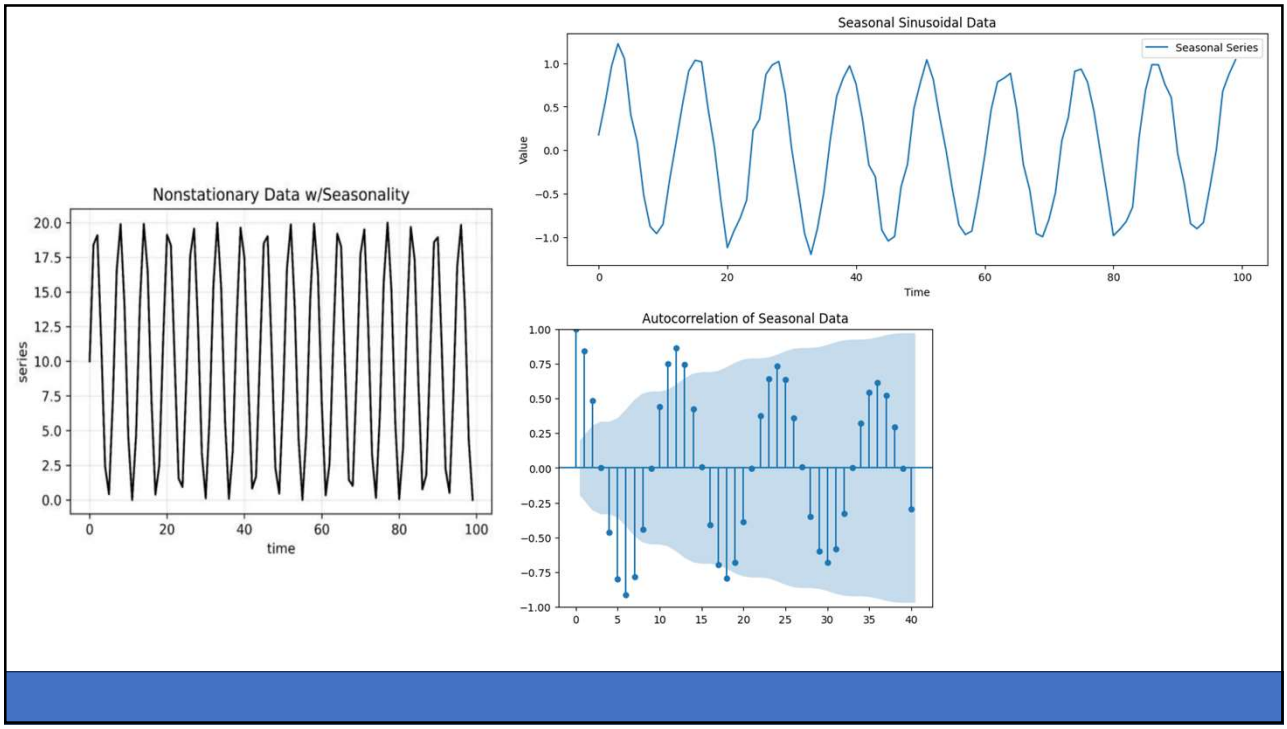
39


Nonstationary Data w/Heteroscedasticity
series
time
20
10
0
−10
−20
−30
0
20
40
60
80
100

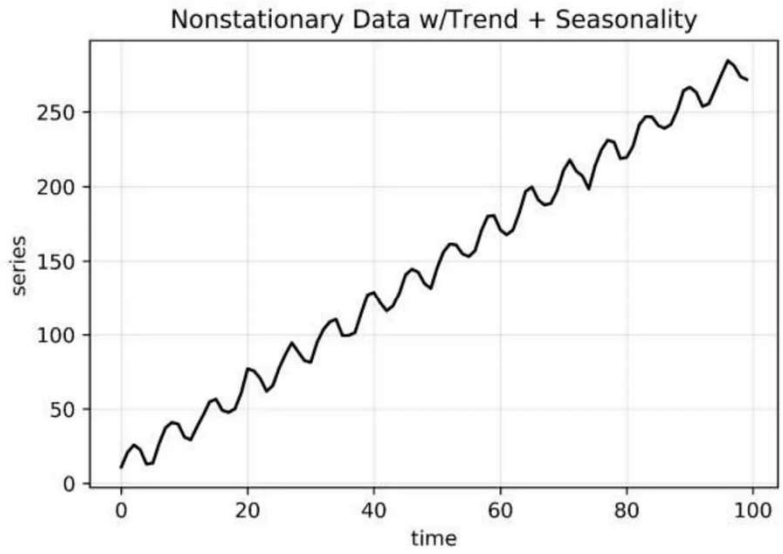
40


Nonstationary Data w/Lagged Structure
series
time
20
15
10
5
0
−5
0
20
40
60
80
100

41


Nonstationary Data w/Seasonality
series
time
Seasonal Sinusoidal Data
Seasonal Series
Value
Time
Autocorrelation of Seasonal Data

42



43

Identifying Nonstationarity

44

How to Identify Nonstationary Time-Series Data

There are several ways to identify nonstationary time-series data:

- Run-sequence plots
- Summary statistics
- Histogram plot
- Augmented Dickey-Fuller test

45

Run-Sequence Plot

A run-sequence plot is simply a plot of your time-series data.

- This should always be your first step in time-series analysis.
- It often shows whether there is underlying structure.
- Be on the lookout for trend, seasonality, and autocorrelation.
- The previous plots are great examples.

46

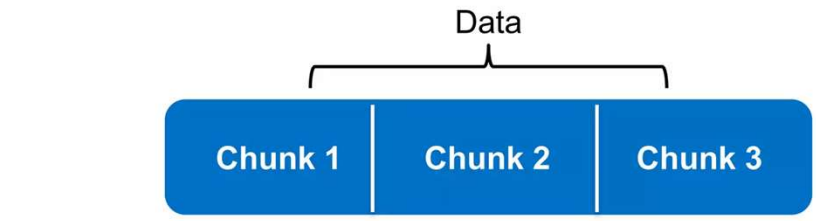
Summary Statistics

Calculating the mean and variance over time is a useful way to discern whether the series is stationary.

- A simple but effective way to do this is to split your data into chunks over time and compute statistics for each chunk.
- Large deviations in either the mean or the variance among chunks are problematic and mean that your data is nonstationary.

47

Stationary

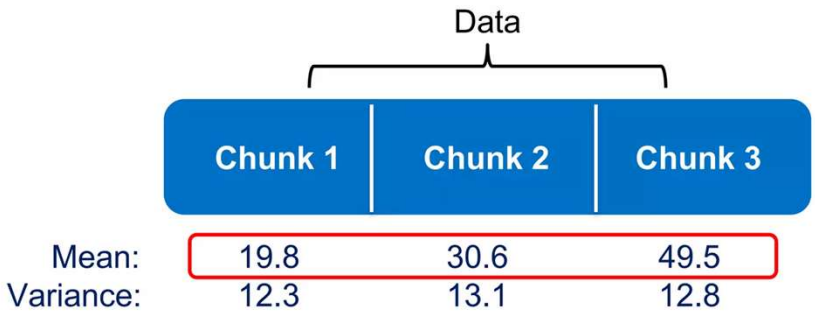


	Chunk 1	Chunk 2	Chunk 3
Mean:	19.8	18.6	18.5
Variance:	12.3	13.1	12.8f

The chunks have similar mean and variance.

48

Nonstationary

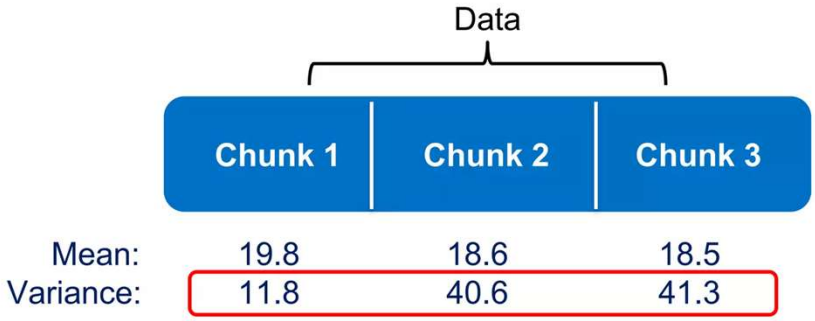


	Chunk 1	Chunk 2	Chunk 3
Mean:	19.8	30.6	49.5
Variance:	12.3	13.1	12.8

There are large deviations in the mean between chunks.

49

Nonstationary



	Chunk 1	Chunk 2	Chunk 3
Mean:	19.8	18.6	18.5
Variance:	11.8	40.6	41.3

There are large deviations in the variance between chunks.

50

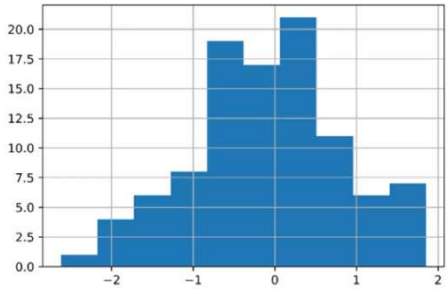
Histogram Plot

A histogram plot gives important clues into a time series' underlying structure.

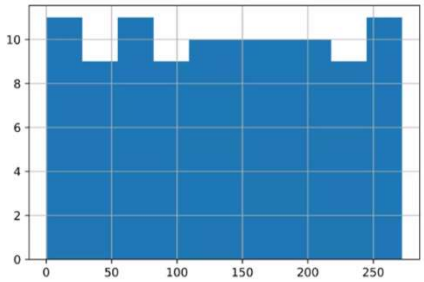
- If you see a distribution that is approximately normal, that's a good indication your time series is stationary.
- If you see a nonnormal distribution, that's a good indication your time series is nonstationary.

51

Stationary



Nonstationary



52

Source	Description
Trend	Long-term increase or decrease
Cyclicality	Economic/business cycles (irregular cycles)
Seasonality	Calendar-based patterns (monthly/quarterly)
Structural Breaks	Events causing sudden level/variance shifts
Unit Roots	Random Walk behavior

53

Unit Root in Time Series

A time series has a unit root if its value at time t is highly dependent on its previous value, making shocks to the series permanent.

Mathematics:

$$X_t = \rho X_{t-1} + \epsilon_t$$

- If $\rho = 1$ → **unit root** → random walk (non-stationary).
- If $|\rho| < 1$ → stationary process.

Key Features:

- Non-stationary mean & variance.
- Autocorrelation decays very slowly.
- Past shocks have lasting effects.

Importance

- Many statistical models require stationarity.
- Unit root processes can cause spurious regression results.

Testing:

- Augmented Dickey-Fuller (ADF) test
- KPSS test

Solution: Differencing: $\Delta Xt = Xt - Xt - 1$ converts many unit root series into stationary series.

54

Basic Dickey-Fuller Setup

- Equation: $y_t = \mu + \phi_1 y_{t-1} + \epsilon_t$
- Hypotheses:
  - $H_0 : \phi_1 = 1$ → Unit root (non-stationary)
  - $H_1 : \phi_1 < 1$ → No unit root (stationary)

Rewritten form for testing

- $y_t - y_{t-1} = \mu + (\phi_1 - 1) y_{t-1} + \epsilon_t$
- Let $\delta = \phi_1 - 1$, so $\Delta y_t = \mu + \delta y_{t-1} + \epsilon_t$

Test Statistic

$$t_{\hat{\delta}} = \frac{\hat{\delta}}{\text{se}(\hat{\delta})}$$

$\text{se}(\hat{\delta})$ is the standard error of $\hat{\delta}$.

Decision Rule

1. Compare $t_{\hat{\delta}}$ with the **Dickey-Fuller critical value** (DF_{critical}).
2. If:
   - $t_{\hat{\delta}} < DF_{\text{critical}}$ → **Reject** H_0 (stationary).
   - $t_{\hat{\delta}} > DF_{\text{critical}}$ → **Do not reject** H_0 (non-stationary).

55

ADF Extension (Augmented Dickey-Fuller)

- Includes lagged differences to account for higher-order correlation:

$$\Delta y_t = \mu + \delta y_{t-1} + \sum_{i=1}^{p} \beta_i \Delta y_{t-i} + \epsilon_t$$

- Hypotheses for ADF:
  - $H_0 : \delta = 0$ → Unit root
  - $H_1 : \delta < 0$ → Stationary

Test Statistic

- $t_{\hat{\beta}_i} = \frac{\hat{\beta}_i}{Se(\hat{\beta}_i)}$

56

How to Test for Stationarity | **Augmented Dickey-Fuller (ADF) Test** | H_o: Series has a unit root (non-stationary) | If p-value < 0.05 → reject H_o → likely stationary | Interpretation of ADF Statistic and critical values

Augmented Dickey-Fuller Test

The Augmented Dickey-Fuller test is a hypothesis test that tests specifically for stationarity.

- We generally say that the series is nonstationary if the p-value is less than 0.05.
- It is a less appropriate test to use with small datasets or when heteroscedasticity is present.
- It is best to pair ADF with other techniques, such as run-sequence plots, summary statistics, or histograms.

57

Stationary Time Series (mean, variance, autocorrelation constant over time)

1.Temperature anomalies (after seasonal adjustment)
1. Example: Daily temperature anomalies (i.e., deviations from long-term average) in a city.
2. Why: After removing seasonality, the fluctuations around the mean tend to be stable.

2.White noise series
1. Example: Sensor noise in electronic equipment.
2. Why: Purely random fluctuations with constant statistical properties.

3.Financial returns (not prices)
1. Example: Daily returns of a stock (percentage change, not price).
2. Why: Mean return and variance often remain constant over short time frames.

4.Machine vibration data under stable operation
1. Example: Sensor readings from a turbine running at constant load.
2. Why: No trend or seasonal component; variance remains constant if the machine is healthy.

Non-Stationary Time Series (mean or variance changes over time)

1.Global average temperature
1. Example: Yearly average global surface temperature.
2. Why: Long-term upward trend due to climate change.

2.Population growth
1. Example: India's population over decades.
2. Why: Shows exponential or logistic growth—clearly trending.

3. Stock prices
1. Example: Daily closing prices of Tesla or Nifty 50.
2. Why: Clear trends, volatility clustering, and unpredictable shocks.

4. GDP of a country
1. Example: Quarterly GDP of the US or India.
2. Why: Strong upward trend due to economic growth, subject to shocks and cycles.

58

Concept	Definition	Stationary?	Memory/Dependence	Example
Random Numbers	Individual draws from some probability distribution (uniform, normal, etc.).	Not a time series (depends on distribution).	Independent (if generated properly).	rand() in Python
White Noise	Sequence of random numbers over time with mean = 0, constant variance, no autocorrelation.	Stationary	No memory	Measurement errors
Random Walk	Cumulative sum of white noise. Current value = past value + random shock.	Non-stationary (variance grows over time).	Strong memory (depends on past).	Stock prices
Brownian Motion (Wiener Process)	Continuous-time limit of a random walk.	Non-stationary (variance $\propto$ time).	Strong memory	Particle motion in fluid
AR(p) Process	Autoregressive: each value depends on previous values + white noise.	Depends on parameters	Has short-term memory	AR(1): $X_t=\phi X_{t-1}+\varepsilon_t$
MA(q) Process	Moving Average: each value depends on past white noise terms.	Stationary (usually)	Memory of past shocks	$X_t=\varepsilon_t+\theta_1\varepsilon_{t-1}$
ARIMA	Combination of AR + I (integration, i.e., differencing → random walk-like) + MA.	Depends on differencing	Long/short memory	Forecasting models

59

i.i.d. (Independent & Identically Distributed)

Definition
A sequence of random variables is i.i.d. if:
1. Independent → Each observation does not affect the others.
2. Identically Distributed → All observations come from the same probability distribution.

Examples
- Tossing a fair coin → Bernoulli(0.5)
- Rolling a fair die → Uniform(1,6)
- Generating samples from Normal(0,1)

Not i.i.d.
- Time series data (stock prices, temperature)
- Survey with groups from different distributions

Importance
- Foundation of Law of Large Numbers & Central Limit Theorem
- Common assumption in statistics & machine learning

60

Central Limit Theorem (CLT)

The CLT says that the **average of many independent random variables (samples) tends to be normally distributed**, even if the original data is not normal.

Statement
If $X_1, X_2, ..., X_n$ are *i.i.d.* random variables with mean μ and variance σ^2, then the sample mean:

$\bar{X}_n = (1/n) \Sigma X_i$

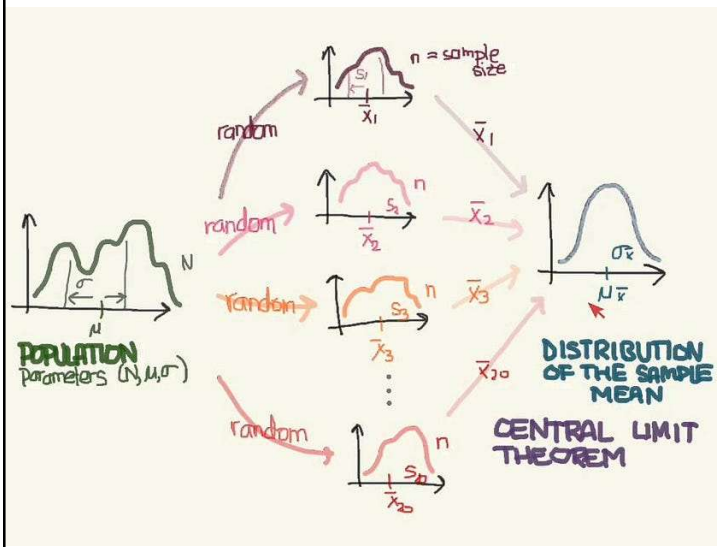
approaches a Normal distribution as $n \to \infty$, regardless of the original distribution.

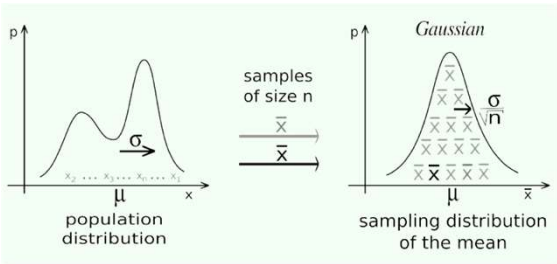
Mathematical Form
$(\bar{X}_n - \mu) / (\sigma/\sqrt{n}) \to N(0,1)$

Intuition
- Averages of many independent samples look Normal (bell-shaped).
- Larger n → closer to Normal distribution.

Importance
- Basis of confidence intervals & hypothesis testing.
- Explains why Normal distribution appears everywhere.

61





62

Common Transformations

63

How to Transform Nonstationary Time-Series Data

There are several ways to transform nonstationary time-series data:

- Remove trend (constant mean)
- Remove heteroscedasticity with log (constant variance)
- Remove autocorrelation with differencing (exploit constant structure)
- Remove seasonality (no periodic component)
- Oftentimes you'll have to do several of these on one dataset!

64

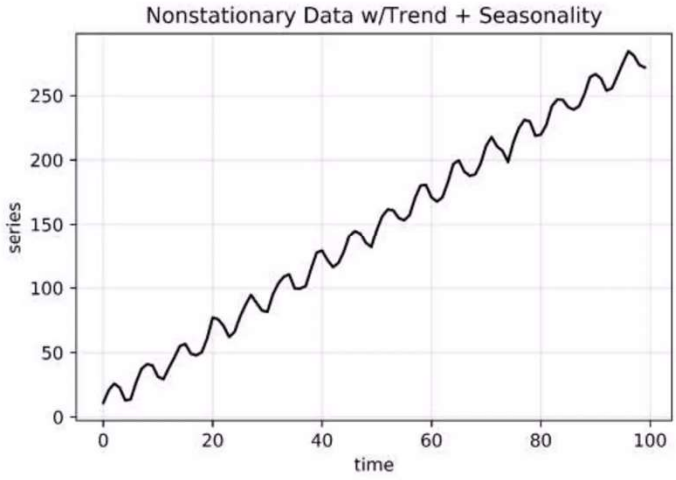
Example 1: Trend & Seasonality Present

A time series with a trend or seasonality component is a nonstationary series. To make it stationary, we can do the following:

- Subtract the trend so that the series has constant mean.
- Subtract the seasonality so that the series has no periodic component.

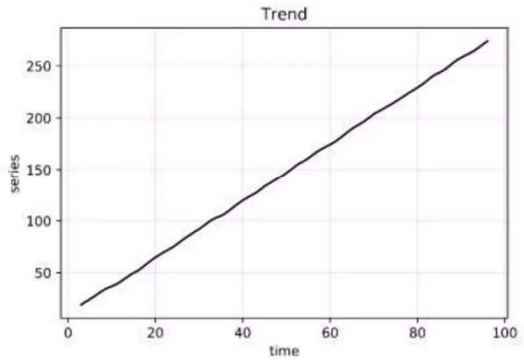
65

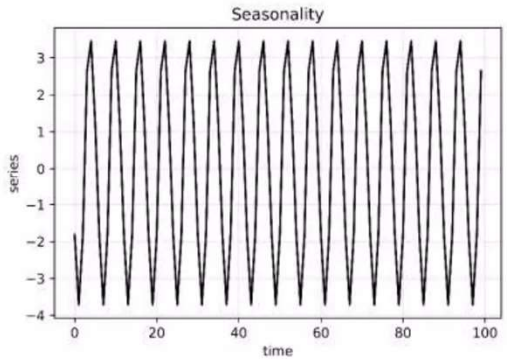
Example 1: Trend & Seasonality Components



66

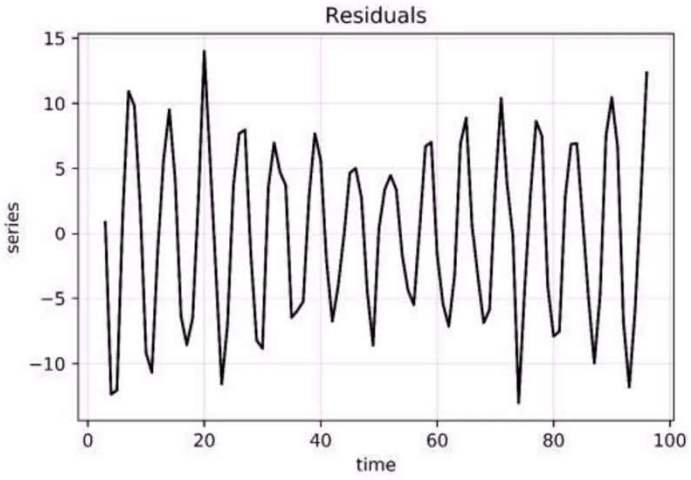
Example 1: Trend & Seasonality Components





67

Example 1: Trend & Seasonality Removed



68

Example 2: Heteroscedasticity Present

A time series with differing variances in two distinct regions is a nonstationary series. To make it stationary, we can do the following:

- Apply the log transformation.
- This squashes the larger values so that the variances are closer.

69

$$r_i = \frac{p_i - p_j}{p_j}$$

$$r_i = \frac{p_i}{p_j} - \frac{p_j}{p_j}$$

$$1 + r_i = \frac{p_i}{p_j}$$

$$log(1 + r_i) = log(\frac{p_i}{p_j})$$

$$log(1 + r_i) = log(p_i) - log(p_j)$$

70

Price and Return Series

Daily price data for General Electric, for time span 4787 trading days over a period of 1982- 2000.

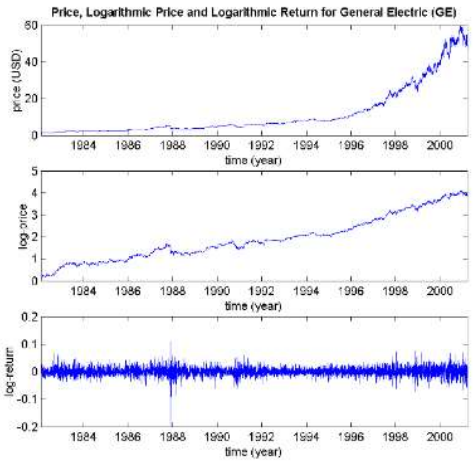
Daily closure price of GE: $P_{GE}(t)$

Daily logarithmic price: $lnP_{GE}(t)$

Daily logarithmic return:

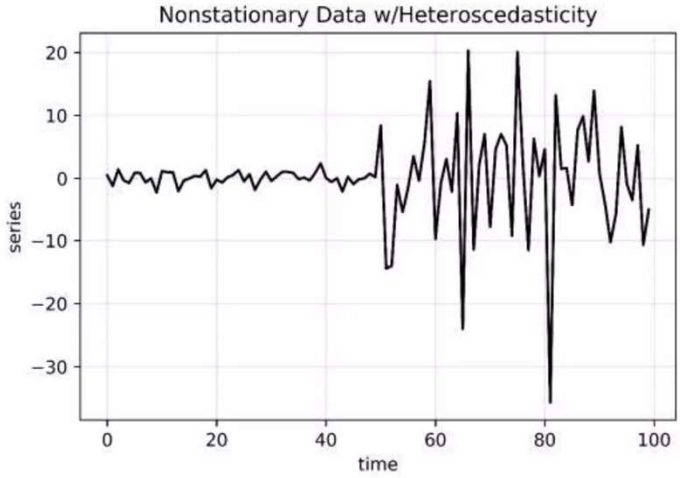
$r_{GE}(t) = lnP_{GE}(t) - lnP_{GE}(t-1)$

$\approx [P_{GE}(t) - P_{GE}(t-1)]/P_{GE}(t)$



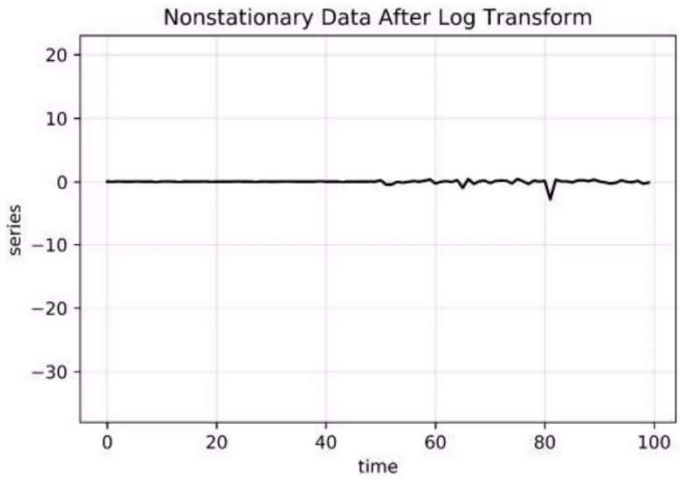
71

Example 2: Heteroscedasticity Present



72

Example 2: Heteroscedasticity Squashed



73

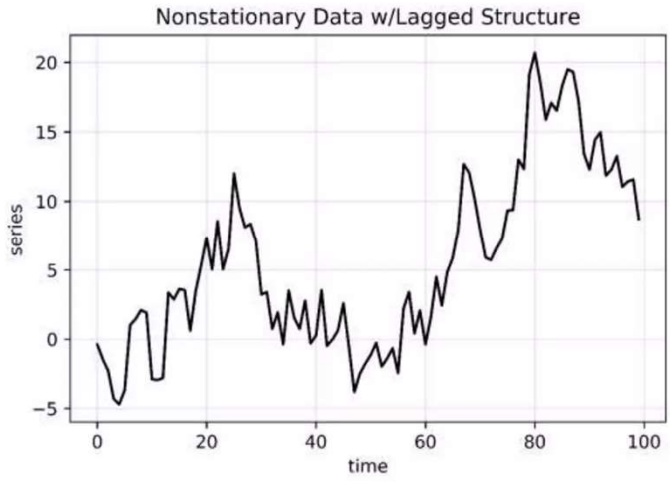
Example 3: Autocorrelation Present

Say a given time series has autocorrelation with a lag of 1. By definition, this is a nonstationary series in its current form. To make it stationary, we can do the following:

- Difference the data by subtracting by a specific lag.
- How you determine the appropriate lag will be covered in the future

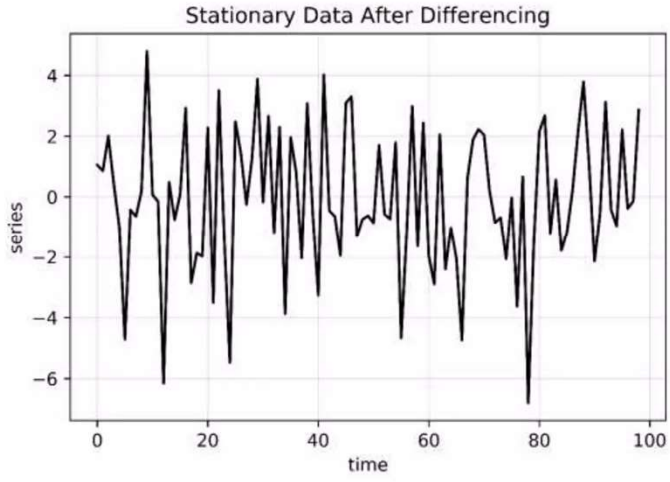
74

Example 2: Autocorrelation Present



75

Example 2: Autocorrelation Removed



76

Auto-**C**orrelation **F**unction (ACF)
vs.
Partial **A**uto-**C**orrelation **F**unction (PACF)

77

ACF (**A**uto-**C**orrelation **F**unction):

The correlation between the observation at the current time spot and the observations at previous time spots.

Definition:

- Measures correlation between X_t and its lagged values X_{t-h}.
- Captures both **direct and indirect** effects of lags.

Formula:

$$\rho(h) = \frac{\mathrm{Cov}(X_t, X_{t-h})}{\mathrm{Var}(X_t)} = \frac{E[(X_t - \mu)(X_{t-h} - \mu)]}{\sigma^2}$$

Key Points:

- ACF shows correlation at **lag** h.
- Useful for identifying **overall correlation pattern**.
- Does not separate direct vs. indirect influence.

78

PACF (**P**artial **A**uto-**C**orrelation **F**unction)

The correlation between observations at two time points, given that we consider both observations are correlated with observations at other time points.

For example, today's stock price can be correlated to the day before yesterday, and yesterday can also be correlated to the day before yesterday. Then, PACF of yesterday is the "real" correlation between today and yesterday after taking out the influence of the day before yesterday.

Definition:

- Measures **pure/direct correlation** between X_t and X_{t-h} after removing effects of intermediate lags.

Formula:

- PACF at lag h = coefficient ϕ_{hh} in AR(h) regression:

$$X_t = \phi_{h1}X_{t-1} + \phi_{h2}X_{t-2} + \cdots + \phi_{hh}X_{t-h} + \epsilon_t$$

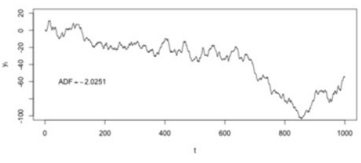
Key Points:

- PACF shows correlation at lag h **net of shorter lags**.
- Cuts off after the **true AR order (p)**.
- Helps determine the **order of AR model**.

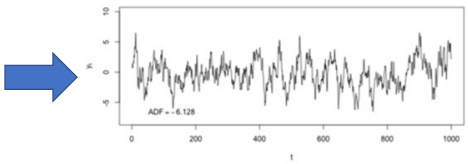
79

Rules of Using ACF and PACF

- Determine if there is an obvious trend in the dataset. If there is, use differencing (i.e., diffing) to "detrend" data. Usually, one-lag differencing is used.

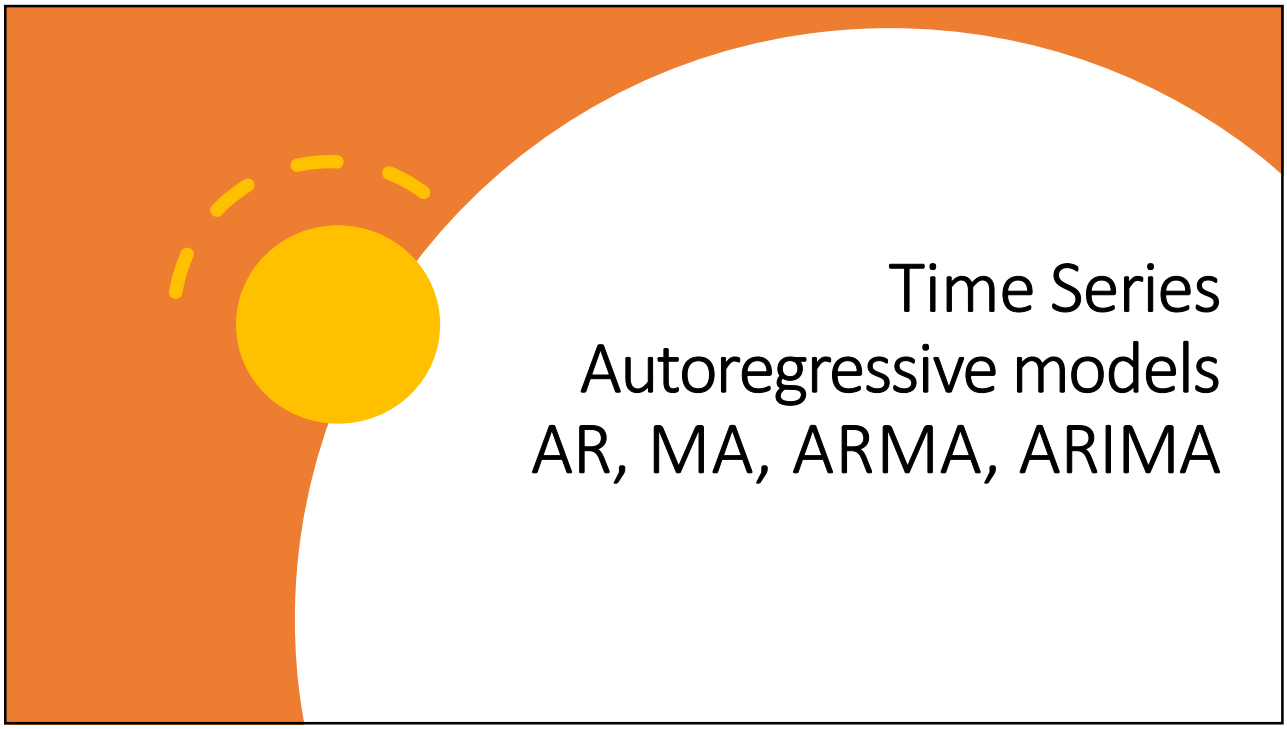


Data with trend



Data after one-lag differencing

80



81

ARIMA Models

ARIMA is an acronym that stands for **A**uto-**R**egressive **I**ntegrated **M**oving **A**verage. Specifically,

- **AR** *Autoregression.* A model that uses the dependent relationship between an observation and some number of lagged observations.
- **I** *Integrated.* The use of differencing of raw observations in order to make the time series stationary.
- **MA** *Moving Average.* A model that uses the dependency between an observation and a residual error from a moving average model applied to lagged observations.

Each of these components are explicitly specified in the model as a parameter.

Note that **AR** and **MA** are two widely used linear models that work on stationary time series, and **I** is a preprocessing procedure to "stationarize" time series if needed.

82

Notations

A standard notation is used of *ARIMA*(*p, d, q*) where the parameters are substituted with integer values to quickly indicate the specific ARIMA model being used.

- **p** The number of lag observations included in the model, also called the lag order.
- **d** The number of times that the raw observations are differenced, also called the degree of differencing.
- **q** The size of the moving average window, also called the order of the moving average.

A value of 0 can be used for a parameter, which indicates to not use that element of the model.

In other words, ARIMA model can be configured to perform the function of an ARMA model, and even a simple AR, I, or MA model.

26 / 77

83

Autoregressive Models

- **Intuition**
  - Autoregressive models are based on the idea that current value of the series, X_t, can be explained as a linear combination of p past values, $X_{t-1}, X_{t-2}, \ldots, X_{t-p}$, together with a random error in the same series.
- **Definition**
  - An autoregressive model of order p, abbreviated $AR(p)$, is of the form

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p} + w_t = \sum_{i=1}^{p} \phi_i X_{t-i} + w_t$$

where X_t is stationary, $w_t \sim wn(0, \sigma_w^2)$, and $\phi_1, \phi_2, \ldots, \phi_p$ ($\phi_p \neq 0$) are model parameters. The hyperparameter p represents the length of the "direct look back" in the series.

84

AR(0) and AR(1)

- The simplest AR process is $AR(0)$, which has no dependence between the terms. In fact, $AR(0)$ is essentially white noise.
- $AR(1)$ can be given by $X_t = \varphi_1 X_{t-1} + \varepsilon_t$.
  - Only the previous term in the process and the noise term contribute to the output.
  - If $|\varphi_1|$ is close to 0, then the process still looks like white noise.
  - If $\varphi_1 < 0$, X_t tends to oscillate between positive and negative values.
  - If $\varphi_1 = 1$ then the process is equivalent to random walk, which is not stationary as the variance is dependent on t (and infinite).

Moving Average Models (MA)

- The name might be **misleading**, but moving average models should not be confused with the moving average smoothing.
- **Motivation**
  - Recall that in AR models, current observation X_t is regressed using the previous observations $X_{t-1}, X_{t-2}, \ldots, X_{t-p}$, plus an error term ε_t at current time point.
  - One problem of AR model is the ignorance of correlated noise structures (which is unobservable) in the time series.
  - In other words, the imperfectly predictable terms in current time, ε_t, and previous steps, $\varepsilon_{t-1}, \varepsilon_{t-2}, \ldots, \varepsilon_{t-q}$, are also informative for predicting observations.

Moving Average Models (MA)

Definition

- A moving average model of order q, or $MA(q)$, is defined to be

$$X_t = \epsilon_t + \theta_1\epsilon_{t-1} + \theta_2\epsilon_{t-2} + \cdots + \theta_q\epsilon_{t-q} = \epsilon_t + \sum_{j=1}^{q}\theta_j\epsilon_{t-j}$$

where $\epsilon_t \sim wn(0, \sigma_\epsilon^2)$, and $\theta_1, \theta_2, \ldots, \theta_q$ $(\theta_q \neq 0)$ are parameters.

- Although it looks like a regression model, the difference is that the ϵ_t is not observable.

- Contrary to AR model, finite MA model is **always stationary**, because the observation is just a weighted moving average over past forecast errors.

87 / 77

87

General MA(q) Process

An important property of *MA*(*q*) models in general is that there are nonzero autocorrelations for the first *q* lags, and $\rho_h = 0$ for all lags $h > q$.

In other words, ACF provides a considerable amount of information about the order of the dependence *q* for *MA*(*q*) process.

Identification of an MA model is often best done with the ACF rather than the PACF.

88

ARMA Models

- Autoregressive and moving average models can be combined together to form **ARMA models**.

Definition

- A time series $\{x_t;\ t = 0, \pm 1, \pm 2, \ldots\}$ is $ARMA(p, q)$ if it is stationary and

$$X_t = \epsilon_t + \sum_{i=1}^{p} \phi_i X_{t-i} + \sum_{j=1}^{q} \theta_j \epsilon_{t-j},$$

where $\phi_p \neq 0$, $\theta_q \neq 0$, and $\sigma_\epsilon^2 > 0$, $\epsilon_t \sim wn(0, \sigma_\epsilon^2)$.

89 / 77

89

Choosing Model Specification

- Recall we have discussed that ACF and PACF can be used for determining ARIMA model hyperparamters *p* and *q*.

	AR(*p*)	*MA*(*q*)	*ARMA*(*p*, *q*)
ACF	Tails off	Cuts off after lag *q*	Tails off
PACF	Cuts off after lag *p*	Tails off	Tails off

- Other criterions can be used for choosing *q* and *q* too, such as AIC (Akaike Information Criterion), AICc (corrected AIC) and BIC (Bayesian Information Criterion).
- Note that the selection for *p* and *q* is not unique.

90

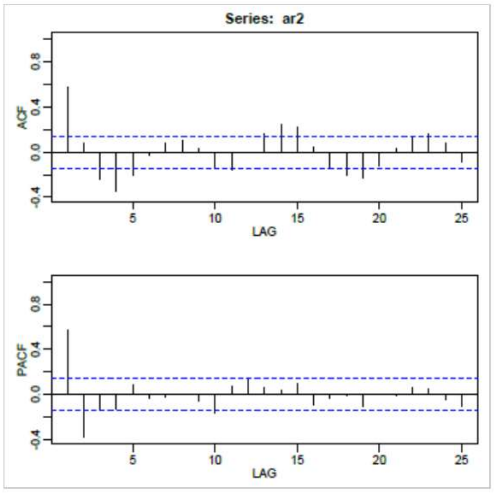
Rules of Using ACF and PACF

- Use PACF to determine the terms used in the AR model. Only significant terms will be chosen. The number of terms determines the order of the model.

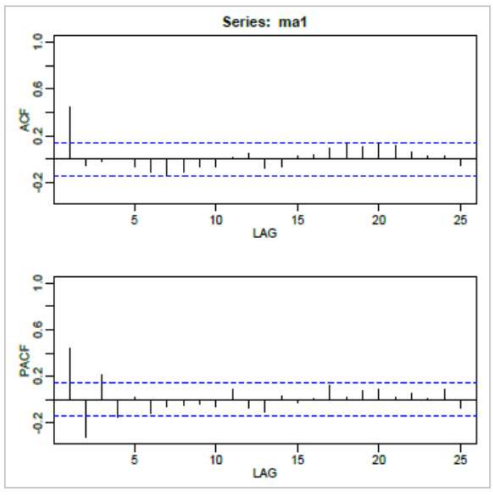
  For example, if the PACF of yesterday's stock price is significant. All other days' PACFs are not significant. Then, yesterday's stock price will be used to predict today's stock price. The AR model is called the first-order autoregression model.
- Use ACF to determine the terms used in the MA model.
- Choose a model by using PACF and ACF charts together.
- Use the "simpler" model if several models could work.

91

Examples

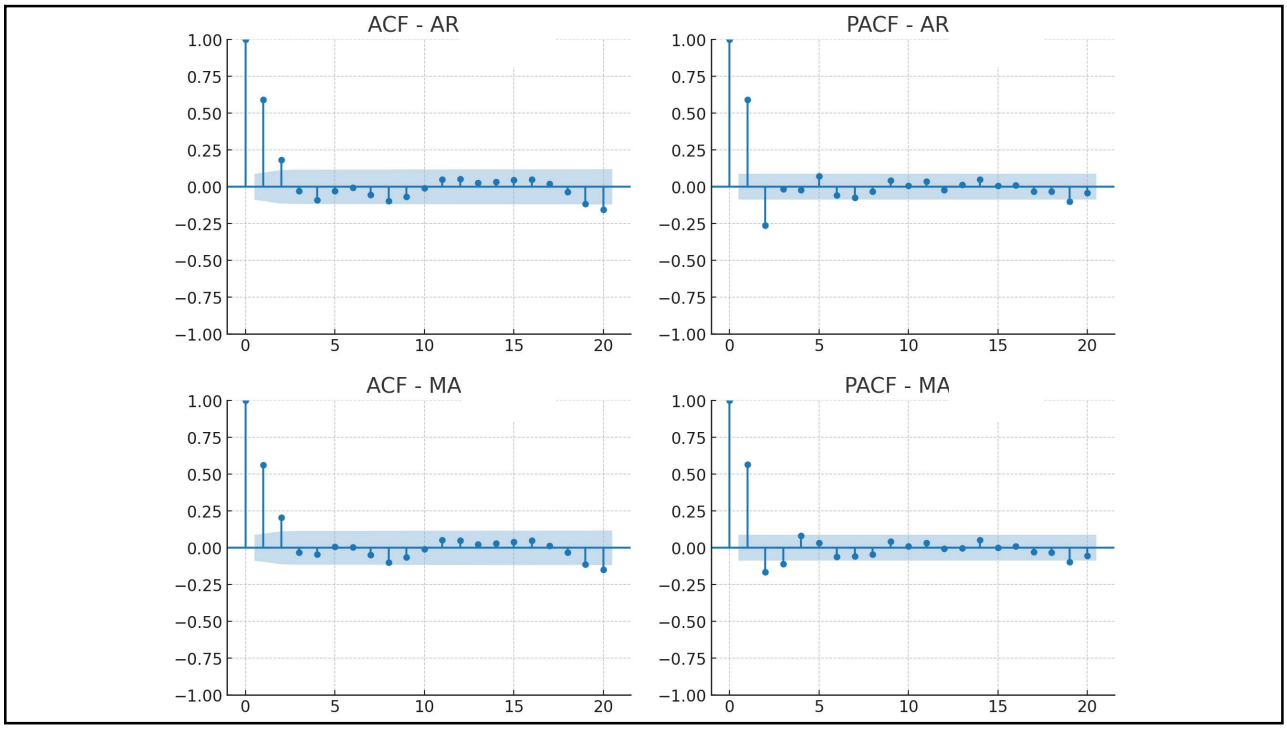


AR(2) Example

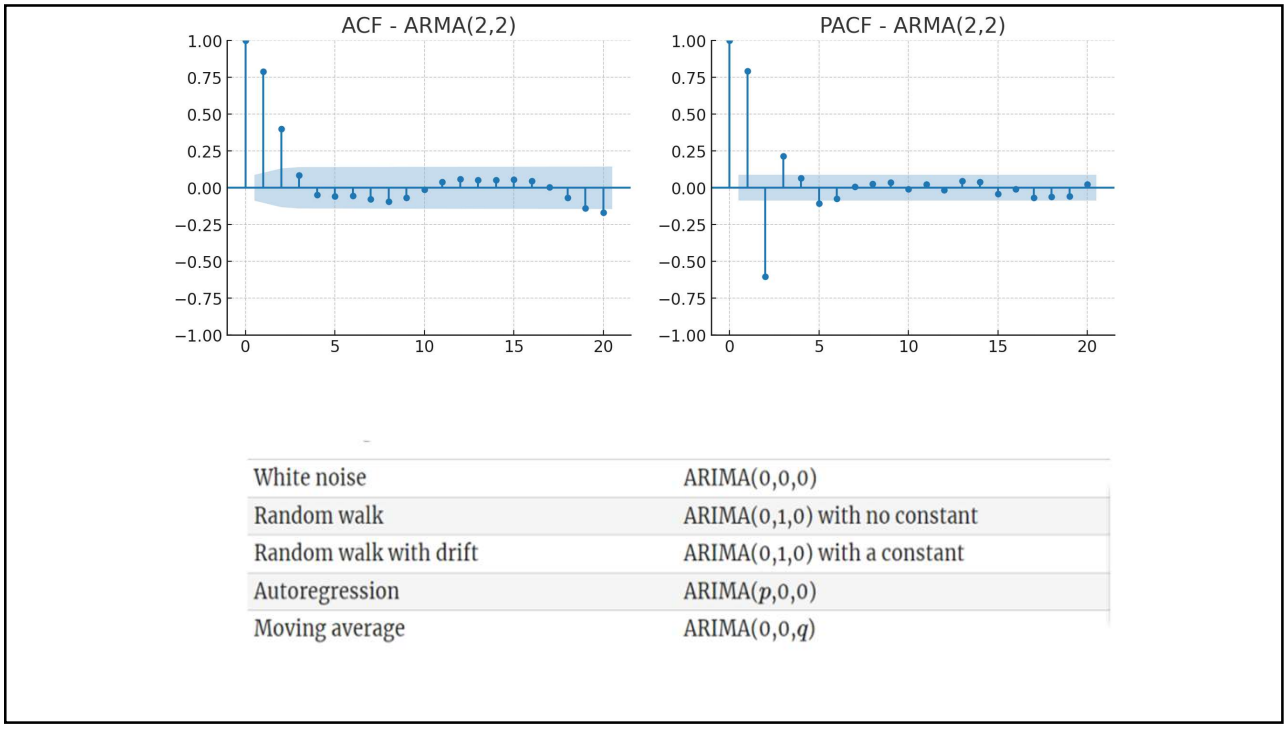


MA(1) Example

92



93



White noise	ARIMA(0,0,0)
Random walk	ARIMA(0,1,0) with no constant
Random walk with drift	ARIMA(0,1,0) with a constant
Autoregression	ARIMA(p,0,0)
Moving average	ARIMA(0,0,q)

94

Two most important extension of ARIMA

1. SARIMA (Seasonal ARIMA)

- Extends ARIMA to handle **seasonality**.
- Model:

$$ARIMA(p,d,q)(P,D,Q)_s$$

where $(P,D,Q)_s$ are the seasonal AR, differencing, and MA terms with period s.

- **Why important:** Time series with repeating patterns (monthly, quarterly, yearly) can't be captured by plain ARIMA, but SARIMA handles it directly.
- **Example:** Monthly sales data with yearly seasonal spikes.

2. ARIMAX (ARIMA with Exogenous Variables)

- Adds external predictors to ARIMA:

$$y_t = \text{ARIMA}(p,d,q) + \beta X_t$$

where X_t = external regressors.

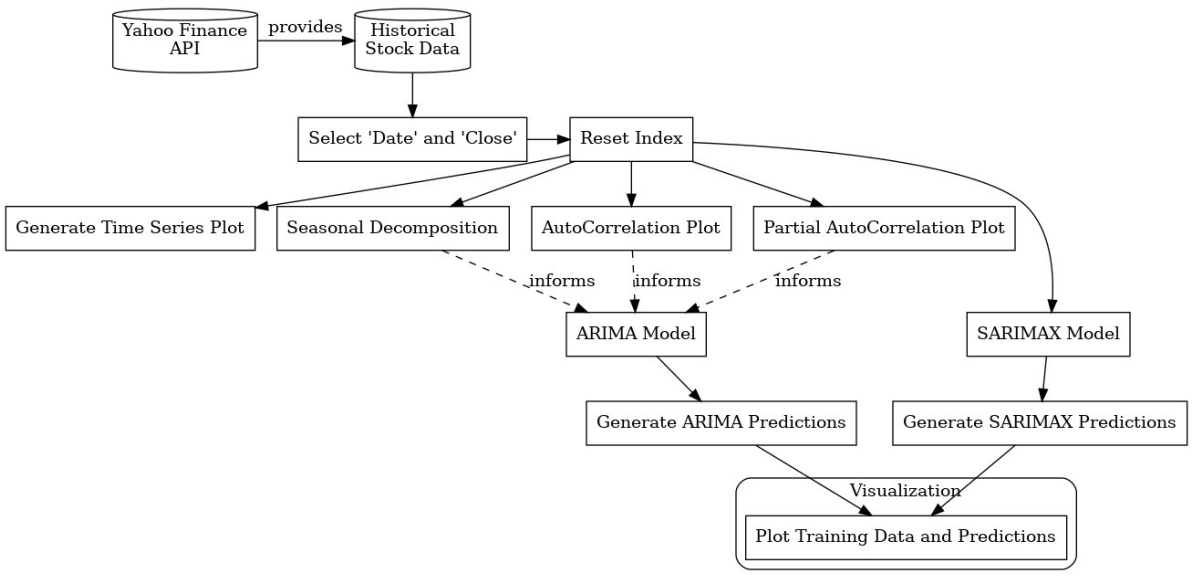
- **Why important:** Many real-world series (like demand, stock returns, macroeconomic data) are influenced by outside factors, not just their own past.
- **Example:** Forecasting electricity demand using ARIMA + temperature as an exogenous variable.

Quick summary:

SARIMA → handles seasonality.

ARIMAX → handles external factors.

95



96

Fat-Tailed Distribution in Stock Returns

1. The Normal Distribution Assumption

Finance theory (e.g., Black–Scholes, CAPM) often assumes:

$$r_t \sim N(\mu, \sigma^2)$$

- Symmetric bell shape
- Thin tails → extreme returns are extremely rare

But real stock returns deviate strongly from this.

2. What Are Fat Tails?

- Fat tails = higher probability of extreme events compared to the normal distribution.
- In practice, this means crashes and spikes happen far more often than Gaussian models predict.

3. Kurtosis and Tail Thickness

Kurtosis measures the **peakedness** and **tail heaviness** of a distribution.

- Formula (for standardized data):

$$K = \frac{E[(X-\mu)^4]}{\sigma^4}$$

- Normal distribution has kurtosis = 3 (this is called **mesokurtic**).

Types of Kurtosis

1. **Mesokurtic (K = 3)**
   - Normal distribution baseline
   - Moderate peak, thin tails
   - Extreme events occur rarely
2. **Leptokurtic (K > 3)**
   - Fat-tailed distribution (typical for stock returns)
   - Sharper peak in the center (many small daily changes)
   - Heavier tails → extreme gains/losses occur more often
   - Student's t-distribution is an example
3. **Platykurtic (K < 3)**
   - Flat-topped, light tails
   - Extreme events less likely than normal
   - Rare in finance

97

4. Stock Returns and Leptokurtosis

- Empirical evidence:
  - Daily, weekly, and even monthly returns show excess kurtosis (K > 3)
  - Small price changes are very frequent (high peak)
  - Market crashes and rallies are more frequent than normal predicts (fat tails)
- Interpretation:
  - Most of the time, returns are calm and clustered around the mean
  - Occasionally, extreme shocks occur → both upside (bubbles) and downside (crashes)

5. Why Important in Finance?

- Leptokurtic distributions explain why models based only on normality underestimate risk
- Extreme movements are not "rare accidents" but built-in features of financial data
- This motivates the use of ARCH/GARCH, Student's t-distributions, and stable laws in finance

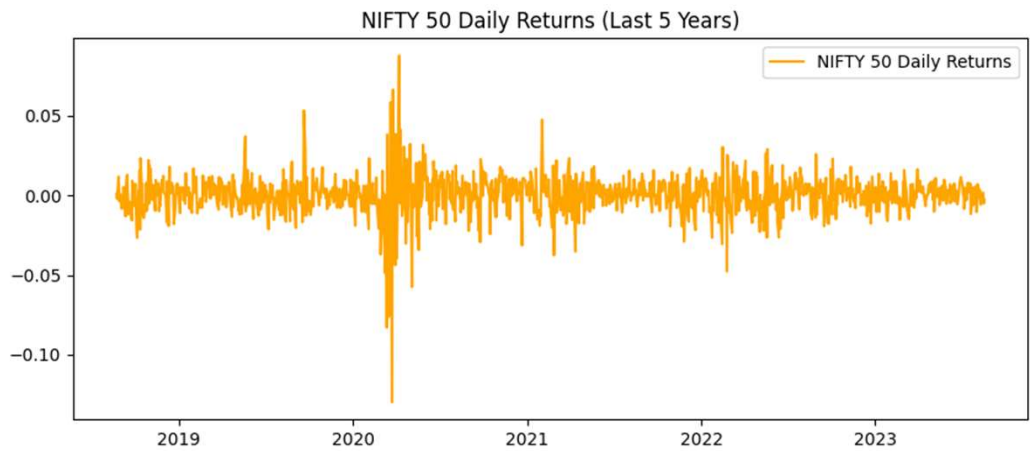
98

Q-Q Plots

Self-study

99

ARCH and GARCH Model
Generalized Autoregressive Conditional Heteroskedasticity



100

Volatility in Financial Markets

Definition

Volatility measures the **degree of variation** in the price of a financial asset (stocks, bonds, currency, etc.) over time.

It reflects the **uncertainty and risk** in the market.

Formula (Standard Deviation of Returns):

$$\sigma = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(r_i - \bar{r})^2}$$

Where r_i = return at time i, $\bar{r}$ = average return, N = number of observations.

Important

- Key input for **risk management** and **derivative pricing**.
- Helps investors assess **market uncertainty**.
- Higher volatility = Higher risk & potential returns.

Measurement

- Standard deviation of returns/innovations.
- Models: **ARCH/GARCH** for time-varying volatility.

101

ARCH and GARCH Model: Returns Dependence

Models:

- ARCH: autoregressive conditional heteroscedasticity
- GARCH: generalized autoregressive conditional heteroscedasticity

Constant vs. Non-constant Variance

- **Unconditional variance** (long-run variance):

$$Var(u_t)$$

is constant — the process as a whole doesn't "blow up" or shrink forever.

- **Conditional variance** (variance given the past):

$$Var(u_t \mid u_{t-1})$$

is **time-varying** — it depends on recent shocks and past variances.

- If there was a big shock yesterday, today's variance will likely be larger.
- This is what creates **volatility clustering**.

Note:

- Even though **returns are uncorrelated** (they look like white noise),
- Their **second moments (variances)** show clear dependence.
- ARCH/GARCH models use this dependence to forecast **next-period volatility**.
- The main idea behind the ARCH/GARCH model is that the log-returns rt are usually uncorrelated but there is still dependence (**squared returns (volatility)**).

102

1. Basic Volatility and Variance Rate Definitions

- Returns are measured as:
  - **Log return:**

$$r_t = \ln\left(\frac{S_t}{S_{t-1}}\right)$$

  - **Simple return:**

$$r_t = \frac{S_t - S_{t-1}}{S_{t-1}}$$

- Volatility (variance of returns) often estimated with rolling window:

$$\sigma_n^2 = \frac{1}{m}\sum_{i=1}^{m} r_{n-i}^2$$

(assuming mean ≈ 0).

2. The Mean Process

- Observed return is split into mean + residual:

$$r_t = \mu_t + u_t$$

where

- r_t: observed return at time t
- μ_t: mean process (constant or ARMA)
- u_t: residual (error term).
  Innovation

103

3. Error Terms (Residuals)

- Defined as:

$$u_t = r_t - \mu_t$$

- This is the **unexpected part of returns**.
- These u_t are the inputs for ARCH/GARCH volatility modeling.

4. Residuals and White Noise

- Residuals are modeled as:

$$u_t = \sigma_t \epsilon_t$$

- Where:
  - ϵ_t: white noise, i.i.d. with mean 0, variance 1
  - σ_t: conditional standard deviation

u_t = **scaled white noise**, hence variance of residuals changes over time.

5. ARCH Model (Engle, 1982)

- Conditional variance depends on past squared residuals:

$$\sigma_t^2 = \alpha_0 + \alpha_1 u_{t-1}^2 + \alpha_2 u_{t-2}^2 + \cdots$$

6. GARCH Model (Bollerslev, 1986)

- Extends ARCH by adding persistence:

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

- Interpretation:
  - α: effect of past shocks
  - β: persistence of volatility

104

Volatility: Weighting Scheme

We want to give more weights to recent data, hence equation

$$\sigma_n^2 = \sum_{i=1}^{m} \alpha_i \, u_{n-i}^2$$

where

- α is positive: $\alpha > 0$
- less weight is given to older observations: $\alpha_i < \alpha_j$ when $i > j$
- sum to unity: $\sum_{i=1}^{m} \alpha_i = 1$

Further assume that there is a long-run average variance rate

$$\sigma_n^2 = \gamma V_L + \sum_{i=1}^{m} \alpha_i \, u_{n-i}^2 \qquad \text{with} \quad \gamma + \sum_{i=1}^{m} \alpha_i = 1$$

105

The ARCH(m) Model

The *Autoregressive Conditional Heteroskedasticity* (ARCH) model was first developed by Engle in 1982.

The estimate of the variance is based on a long-run average variance and m observations. The older an observation, the less weight it is given.

ARCH(1): $\sigma_t^2 = \omega + \alpha_1 \, u_{t-1}^2$ where $\omega = \gamma V_L$

Generalizing:

$$\begin{cases} u_t = \sigma_t \epsilon_t \\ \sigma_t^2 = \omega + \underbrace{\sum_{i=1}^{p} \alpha_i u_{t-i}^2}_{\text{ARCH term}} \end{cases}$$

106

The GARCH(p, q) Model

The *Generalized Autoregressive Conditional Heteroskedasticity* (GARCH) model was first introduced by Bollerslev in 1986.

The simplest version of the model is the GARCH(1,1) one, where the variance rate is calculated from a long-run average variance rate, V_L, as well as from σ_{n-1} and u_{n-1}. Defined as:

$$\sigma_t^2 = \gamma V_L + \alpha_1 u_{t-1}^2 + \beta_1 \sigma_{t-1}^2 = \omega + \alpha_1 u_{t-1}^2 + \beta_1 \sigma_{t-1}^2$$

where $\alpha_1 + \beta_1 < 1$ in order to ensure stability of the process.

A special case of the GARCH(1,1) model is the Exponentially weighted moving average (EWMA) model, where $\gamma = 0$, $\alpha_1 = 1 - \lambda$ and $\beta_1 = \lambda$.

$$\sigma_t^2 = \lambda \sigma_{t-1}^2 + (1 - \lambda) u_{t-1}^2$$

107

The GARCH(p, q) Model (Cont.)

The *Generalized Autoregressive Conditional Heteroscedasticity* (GARCH(p,q)) model is defined by the following system of equations:

$$\begin{cases} u_t = \sigma_t \epsilon_t \\ \sigma_t^2 = \omega + \underbrace{\sum_{i=1}^{p} \alpha_i u_{t-i}^2}_{\text{ARCH term}} + \underbrace{\sum_{j=1}^{q} \beta_j \sigma_{t-j}^2}_{\text{GARCH term}} \end{cases}$$

where $\omega > 0$, $\alpha_i \geq 0$ and $\beta_j \geq 0$ and $\alpha_i + \beta_i < 1$ in order to ensure the finiteness of the unconditional variance.

How can we estimate ω, α and β? $\Rightarrow$ Maximum Likelihood Estimation

Basically, choosing values for the parameters that maximize the chance (or likelihood) of the data occurring.

108

Autocorrelation in Returns

- There is usually a certain form of heteroskedasticity in a series of returns.
- High volatility today can lead to high volatility tomorrow.
- Variances today and tomorrow are somehow related.
- This form of heteroskedasticity implies that there will be autocorrelation in squared returns. → ARCH Effect

109

Two most important extensions

1. EGARCH (Exponential GARCH – Nelson, 1991)

- Model volatility in logs:

$$\ln(\sigma_t^2) = \alpha_0 + \beta \ln(\sigma_{t-1}^2) + \gamma \frac{u_{t-1}}{\sigma_{t-1}} + \theta \left(\left| \frac{u_{t-1}}{\sigma_{t-1}} \right| - E \left| \frac{u_{t-1}}{\sigma_{t-1}} \right| \right)$$

- **Why important:**
  - Guarantees positivity of variance automatically.
  - Captures **asymmetry/leverage effect** (bad news raises volatility more than good news).

2. GJR-GARCH / Threshold GARCH (Glosten-Jagannathan-Runkle, 1993)

- Conditional variance:

$$\sigma_t^2 = \alpha_0 + \alpha u_{t-1}^2 + \gamma I_{t-1} u_{t-1}^2 + \beta \sigma_{t-1}^2$$

where $I_{t-1} = 1$ if $u_{t-1} < 0$, else 0.

- **Why important:**
  - Simple way to allow **different impact of positive vs negative shocks.**
  - If $\gamma > 0$, volatility reacts more to negative returns (bad news).

So in short:

EGARCH → log variance, asymmetric, leverage effect.

GJR-GARCH → threshold model, negative shocks hit harder.

110

Thank you

111