# Empirical Finance: Methods & Applications

## Introduction & Basic Concepts

Pasquale Della Corte
Imperial College London

Week 1

# Why is Empirical Finance Important?

*"Theorists develop models with testable predictions; empirical researchers document 'puzzles' – stylized facts that fail to fit established theories – and this stimulates the development of new theories. Such a process is part of the normal development of any science."*

John Y. Campbell (2000). *Asset Pricing at the Millemium*, Journal of Finance.

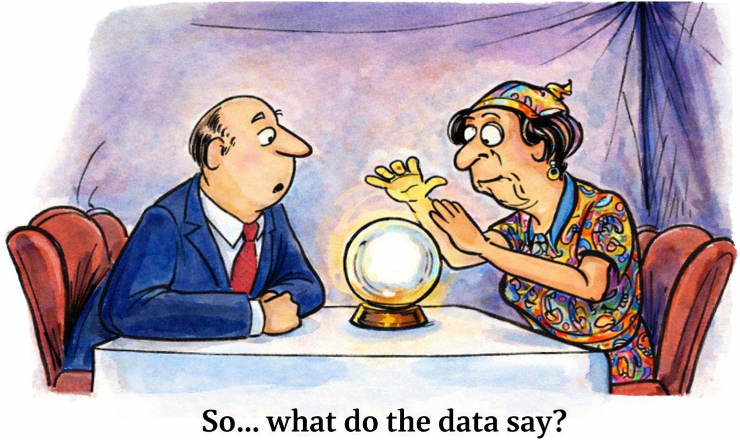
# Why is Empirical Finance Complex?

Finance is not an experimental science and we face difficult problems of inference.

*"The economic world is extremely complicated. There are millions of people and firms, thousands of prices and industries. One possible way of figuring out economic laws in such a setting is by controlled experiments ... like those done by chemists, physicists, and biologists ... Economists have no such luxury when testing economic laws. They cannot perform the controlled experiments of chemists or biologists because they cannot easily control other important factors. Just like astronomers or meteorologists, they usually have to solely use their observation."*

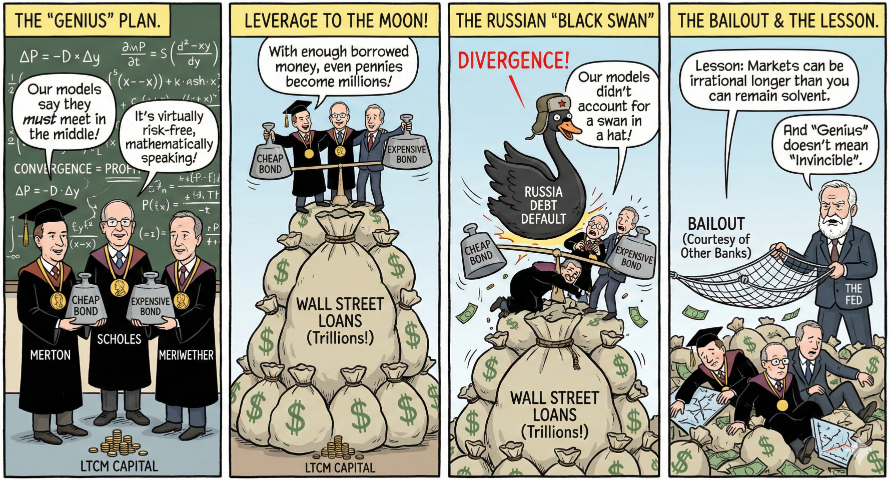
Paul Samuelson & William Nordhaus (1985). *Economics*, McGraw Hill.

# Finance Is Not Just Science — It's Also Art



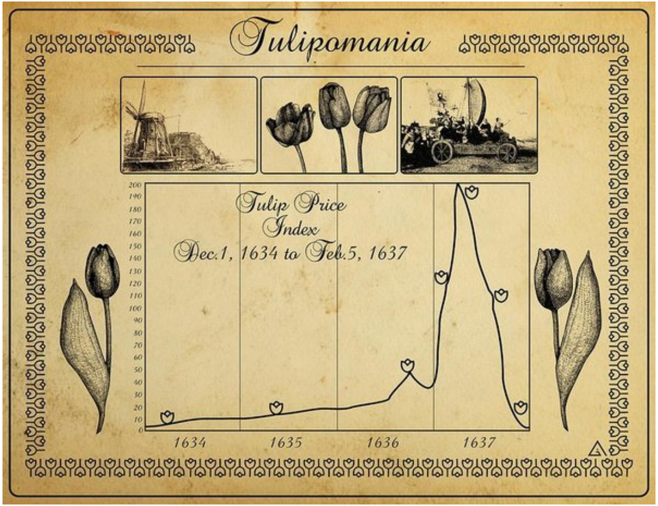

Financial markets often break model assumptions, so knowing when to trust or adjust a model takes intuition and experience.

# When Models Fail: The Collapse of LTCM



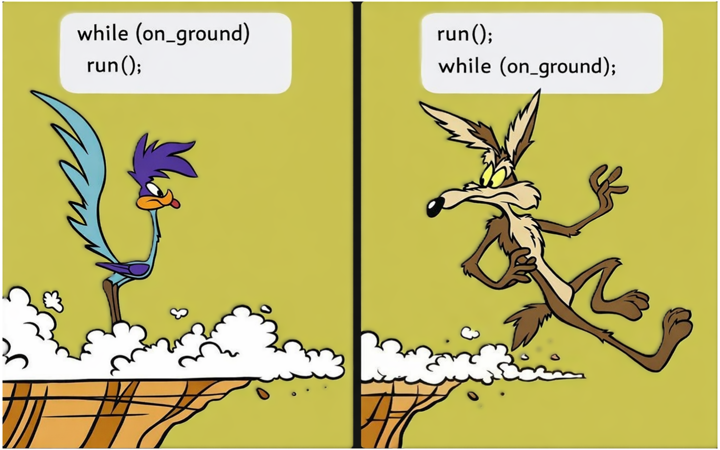

LTCM didn't fail because of bad math: Reality broke the assumptions.

# Empirics Without Models: The Tulip Bubble




Tulip Mania shows what happens when investing without a disciplined model

# The Devil is Always in the Details




Same idea. One tiny detail. Completely different outcome!

Let's Get Started . . .

# What is an Asset Return?

The return of an asset is an important concept of modern finance

- It is the benefit that an investor receives for holding an asset,
- It depends on the holding period of the asset (e.g., one day, one month, etc).

To calculate an asset return, we need the price of the asset

- Price are formed in different markets: auction markets, dealer markets, etc.
- We may observe indicative quotes, effective quotes, or transaction prices.

We can work with simple or log returns

# Simple (or Discrete) Returns

# Simple Return: One Period

$P_t$ is the price of an asset that pays no dividends at time $t$

- Suppose you hold the asset for a single period between times $t$ and $t+1$,

The gross simple return between times $t$ and $t+1$ is calculated as

I could have used $R_{t,t+1}$ but we omit the first $t$ for simplicity

$$R_{t+1} = \frac{P_{t+1}}{P_t}$$

Why t+1 rather than t? Because the return is known ex-post at t+1

The net simple return between times $t$ and $t+1$ is calculated as

I am not using $r_{t+1}$ to avoid confusion with log returns

$$R_{t+1} - 1 = \frac{P_{t+1}}{P_t} - 1$$

We will see log returns later

# Simple Return: Two Periods

The gross simple return between times $t$ and $t+2$ is calculated as

$$R_{t,t+2} = \underbrace{\frac{P_{t+2}}{P_{t+1}}}_{\text{one-period gross return } R_{t+2}} \times \underbrace{\frac{P_{t+1}}{P_t}}_{\text{one-period gross return } R_{t+1}}$$

or alternatively as the product of one-period returns

$$R_{t,t+2} = R_{t+2} \times R_{t+1}$$

Common to drop the first subscript t
only for one-period returns

# Simple Return: Two Periods

The net simple return between times $t$ and $t+2$ is then obtained as

$$R_{t,t+2} - 1 = (R_{t+2} \times R_{t+1}) - 1$$

Recall that for $\varepsilon$ and $\zeta$ near to zero ('small')

$$(1+\varepsilon)(1+\zeta) \approx 1 + \varepsilon + \zeta \longrightarrow \text{we drop the term } \zeta\varepsilon$$

If one-period net returns are small, the two-period net simple return can be written as

$$R_{t,t+2} - 1 \approx (R_{t+2} - 1) + (R_{t+1} - 1)$$

# Simple Return: Multiple Periods

The gross simple return between times $t$ and $t+k$ is calculated as

$$R_{t,t+k} = \frac{P_{t+k}}{P_t} = \underbrace{\frac{P_{t+k}}{P_{t+k-1}}}_{\text{one-period gross return } R_{t+k-1,t+k}} \times \underbrace{\frac{P_{t+k-1}}{P_{t+k-2}}}_{\text{one-period gross return } R_{t+k-1,t+k-2}} \times \ldots \times \underbrace{\frac{P_{t+1}}{P_t}}_{\text{one-period gross return } R_{t+1}}$$

or alternatively as the product of one-period net returns

$$R_{t,t+k} = R_{t+k} \times R_{t+k-1} \times \ldots \times R_{t+1} = \prod_{j=1}^{k} R_{t+j}$$

# Simple Return: Multiple Periods

The net simple return between times $t$ and $t+k$ is then obtained as

$$R_{t,t+k} - 1 = \left[\prod_{j=1}^{k} R_{t+j}\right] - 1$$

or approximately as the sum of one-period returns

$$R_{t,t+k} - 1 \approx \sum_{j=1}^{k} (R_{t+j} - 1)$$

# An Example: S&P500 Index

| Time | Date | S&P500 | Gross Return (Formula) | Gross Return | Net Return (Formula) | Net Return |
|---|---|---|---|---|---|---|
| $t$ | 2024-11-29 | 6032.38 | | | | |
| $t+1$ | 2024-12-02 | 6047.15 | $\frac{6047.15}{6032.38}$ | 1.0024 | $\frac{6047.15}{6032.38}-1$ | 0.0024 |
| $t+2$ | 2024-12-03 | 6049.88 | $\frac{6049.88}{6047.15}$ | 1.0005 | $\frac{6049.88}{6047.15}-1$ | 0.0005 |
| $t+3$ | 2024-12-04 | 6086.49 | $\frac{6086.49}{6049.88}$ | 1.0061 | $\frac{6086.49}{6049.88}-1$ | 0.0061 |
| $t+4$ | 2024-12-05 | 6075.11 | $\frac{6075.11}{6086.49}$ | 0.9981 | $\frac{6075.11}{6086.49}-1$ | −0.0019 |
| $t+5$ | 2024-12-06 | 6090.27 | $\frac{6090.27}{6075.11}$ | 1.0025 | $\frac{6090.27}{6075.11}-1$ | 0.0025 |
| $t+6$ | 2024-12-09 | 6052.85 | $\frac{6052.85}{6090.27}$ | 0.9939 | $\frac{6052.85}{6090.27}-1$ | −0.0061 |
| $t+7$ | 2024-12-10 | 6034.91 | $\frac{6034.91}{6052.85}$ | 0.9970 | $\frac{6034.91}{6052.85}-1$ | −0.0030 |
| $t+8$ | 2024-12-11 | 6084.19 | $\frac{6084.19}{6034.91}$ | 1.0082 | $\frac{6084.19}{6034.91}-1$ | 0.0082 |
| $t+9$ | 2024-12-12 | 6051.25 | $\frac{6051.25}{6084.19}$ | 0.9946 | $\frac{6051.25}{6084.19}-1$ | −0.0054 |
| $t+10$ | 2024-12-13 | 6051.09 | $\frac{6051.09}{6051.25}$ | 1.0000 | $\frac{6051.09}{6051.25}-1$ | 0.0000 |

Source: Datastream. Author's calculations.

# An Example: S&P500 Index

The simple return between times $t$ and $t+10$ in percentage

$$R_{t,t+10} - 1 = [(1.0024 \times 1.0005 \times \ldots \times 1.0000) - 1] \times 100 = 0.31\%$$

or simply 31 basis points

The approximation between times $t$ and $t+10$ in percentage

$$R_{t,t+10} - 1 \approx (0.0024 + 0.0005 + \ldots + 0.0000) \times 100 = 0.32\%$$

or simply 32 basis points

# Average Simple Return

What is the average return of an asset held over multiple periods?

- For example, you want to compare assets held over different periods.

We need to calculate the average simple return $\overline{R}$ as follows

$$\prod_{j=1}^{k} \overline{R} = \prod_{j=1}^{k} R_{t+j}$$

one-period constant return
(unobserved)

one-period actual return
(observed)

So that $\overline{R}$ and $R_t$ deliver an identical payoff over the same holding period.

# Average Simple Return

The average gross simple return is thus given by

$$\overline{R} = \left[\prod_{j=1}^{k} R_{t+j}\right]^{1/k} \longrightarrow \text{Geometric Average}$$

The average net simple return follows as

$$\overline{R} - 1 = \left[\prod_{j=1}^{k} R_{t+j}\right]^{1/k} - 1$$

If returns are small, the average simple return is approximately

$$\overline{R} - 1 \approx \frac{1}{k}\sum_{j=1}^{k}(R_{t+j} - 1) \longrightarrow \text{Arithmetic Average}$$

# An Example: S&P500 Index

The average simple return between times $t$ and $t + 10$ in percentage

$$\overline{R} - 1 = \left[(1.0024 \times 1.0005 \times \ldots \times 1.0000)^{\frac{1}{10}} - 1\right] = \boxed{0.031\%}$$

or simply 3.1 basis points per day

The approximation between times $t$ and $t + 10$ in percentage

$$\overline{R} - 1 \approx \frac{1}{10}(0.0024 + 0.0005 + \ldots + 0.0000) \times 100 = \boxed{0.032\%}$$

or simply 3.2 basis points per day

# Simple Returns in Real Terms

In some cases, we need to adjust returns for inflation

- For example, there are different inflation regimes when using long-span data.

Let's define the gross inflation rate as follows

Consumer Price Index at time t+1

$$\Pi_{t+1} = \frac{CPI_{t+1}}{CPI_t}$$

Consumer Price Index at time t

The Inflation rate is computed as $\Pi_{t+1} - 1$

# Simple Returns in Real Terms

The real gross simple return can be calculated as follows

$$R^r_{t+1} = \frac{P_{t+1}/P_t}{CPI_{t+1}/CPI_t}$$

$$= \frac{R_{t+1}}{\Pi_{t+1}}$$

or simply

The real net simple return can be calculated as follows

$$R^r_{t+1} - 1 = \frac{R_{t+1}}{\Pi_{t+1}} - 1$$

$$\approx R_{t+1} - \Pi_{t+1}$$

if the inflation rate is reasonably small

# An Example: S&P500 Index

| Time | Date | S&P500 | Return (Formula) | Return | CPI | Inflation (Formula) | Inflation |
|---|---|---|---|---|---|---|---|
| $t$ | 2022-12-30 | 3839.50 | | | 298.812 | | |
| $t+1$ | 2023-01-31 | 4076.60 | $\frac{4076.60}{3839.50}$ | 1.0618 | 300.356 | $\frac{300.356}{298.812}$ | 1.0052 |
| $t+2$ | 2023-02-28 | 3970.15 | $\frac{3970.15}{4076.60}$ | 0.9739 | 301.509 | $\frac{301.509}{300.356}$ | 1.0038 |
| $t+3$ | 2023-03-31 | 4109.31 | $\frac{4109.31}{3970.15}$ | 1.0351 | 301.744 | $\frac{301.744}{301.509}$ | 1.0008 |
| $t+4$ | 2023-04-28 | 4169.48 | $\frac{4169.48}{4109.31}$ | 1.0146 | 303.032 | $\frac{303.032}{301.744}$ | 1.0043 |
| $t+5$ | 2023-05-31 | 4179.83 | $\frac{4179.83}{4169.48}$ | 1.0025 | 303.365 | $\frac{303.365}{303.032}$ | 1.0011 |
| $t+6$ | 2023-06-30 | 4450.38 | $\frac{4450.38}{4179.83}$ | 1.0647 | 304.003 | $\frac{304.003}{303.365}$ | 1.0021 |
| $t+7$ | 2023-07-31 | 4588.96 | $\frac{4588.96}{4450.38}$ | 1.0311 | 304.628 | $\frac{304.628}{304.003}$ | 1.0021 |
| $t+8$ | 2023-08-31 | 4507.66 | $\frac{4507.66}{4588.96}$ | 0.9823 | 306.187 | $\frac{306.187}{304.628}$ | 1.0051 |
| $t+9$ | 2023-09-29 | 4288.05 | $\frac{4288.05}{4507.66}$ | 0.9513 | 307.288 | $\frac{307.288}{306.187}$ | 1.0036 |
| $t+10$ | 2023-10-31 | 4193.80 | $\frac{4193.80}{4288.05}$ | 0.9780 | 307.531 | $\frac{307.531}{307.288}$ | 1.0008 |
| $t+11$ | 2023-11-30 | 4567.80 | $\frac{4567.80}{4193.80}$ | 1.0892 | 308.024 | $\frac{308.024}{307.531}$ | 1.0016 |
| $t+12$ | 2023-12-29 | 4769.83 | $\frac{4769.83}{4567.80}$ | 1.0442 | 308.742 | $\frac{308.742}{308.024}$ | 1.0023 |

Source: S&P500 from Datastream and US CPI from FRED. Author's calculations.

# An Example: S&P500 Index

The real net simple return in percentage

$$R^r_{t+1} - 1 = \frac{1.0618 - 1.0052}{1.0052} \times 100 = 5.63\%$$

The approximate real net simple return in percentage

$$R^r_{t+1} - 1 \approx (1.0618 - 1.0052) \times 100 = 5.66\%$$

# Simple Riskless Return

The riskless return is the return on an investment with no risk

- It is a theoretical quantity that does not exists,
- We often use short-term government bond rate with almost no risk of default as a proxy.

Which proxy then?

- The US Treasury Bill (or simply T-Bill) rate is an example of short-term riskless rate
- The German Federal Treasury rate is another example of short-term riskless rate.
- One-month and three-month maturity instruments are generally used.

# Simple Riskless Return

Let $B_t$ be the price of a short-term government bond at time $t$

- A zero-coupon bond that expires after one period (e.g., one month),
- The liquidation price at time $t+1$ is known at time $t$ (e.g., $B_{t+1} = 100$).

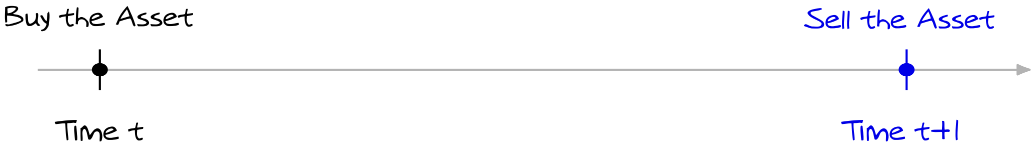
The gross riskless rate between times $t$ and $t+1$ is calculated as

I could have also written $R^f_{t,t+1}$ ← $\boxed{R^f_t} = \frac{B_{t+1}}{B_t}$ → Why t rather than t+1? Because the return is known ex-ante at t

The riskless return between times $t$ and $t+1$ is calculated as

$$R^f_t - 1 = \frac{B_{t+1}}{B_t} - 1$$ → With a one-period bond $B_{t+1} = 1$

# Clarifying Our Notation




Risky asset: The realized return is known at time t+1 $\longrightarrow R_{t+1}$

Riskless asset: The realized return is known at time t $\longrightarrow R_t^f$

# An Example: US Deposit Rates

| USD DEPOSITS | | | | | |
|---|---|---|---|---|---|
| RIC | BID | ASK | Time | High | Low |
| USDOND= | 4.30 | 4.50 | 11:41 | 4.31 | 4.32 |
| USDTND= | 4.25 | 4.32 | 10:00 | 4.32 | 4.32 |
| USDSWD= | 4.27 | 4.61 | 11:50 | 4.35 | 4.40 |
| USD1MD= | 4.28 | 4.53 | 11:41 | 4.41 | 4.39 |
| USD3MD= | 4.34 | 4.59 | 11:41 | 4.43 | 4.31 |
| USD6MD= | 4.55 | 4.65 | 10:00 | 4.55 | 4.44 |
| USD9MD= | 4.36 | 4.61 | 11:41 | 4.59 | 4.45 |
| USD1YD= | 4.50 | 4.56 | 11:49 | 4.62 | 4.56 |

Source: LSEG Workspace

# Fixed Income Instruments

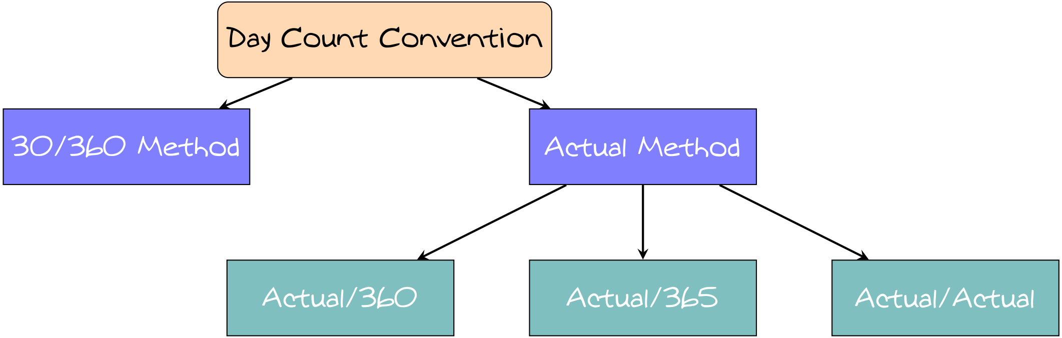
Returns on fixed-income instruments are generally quoted in percentage per annum

- This standardization allows for easy comparison across different maturities,
- The ask quote on the overnight deposit rate is 4.50% per annum,
- The ask quote on the one-month deposit rate is 4.53% per annum.

But we need per-period returns for our calculation

- The actual return for a given period is calculated using the day count convention,
- The day count convention defines how interest accrues over time,
- The day count convention varies by country and maturity.

# Day Count Convention: Some Examples



# Day Count Convention: Some Examples

Most countries generally use Actual/360

$$\text{per-period net return} = \% \text{ per-annum net return} \times \frac{\text{Actual Calendar Days}}{360} \times \frac{1}{100}$$

Japan, the UK, and those historically connected to the UK (e.g., Australia, New Zealand, Hong Kong, Singapore, South Africa) use Actual/365

$$\text{per-period net return} = \% \text{ per-annum net return} \times \frac{\text{Actual Calendar Days}}{365} \times \frac{1}{100}$$

Warning: There are exceptions and you'd better to double check for it!

# Day Count Convention: Some Examples

On 31/10/2024, you invest \$1,000 on a one-month T-Bill with a rate of 5.25%

$$\text{Payoff on } 30/11/2024 = \$1,000 \times \left(1 + \frac{5.25}{100} \times \frac{30}{360}\right) = \$1004.375$$

On 31/10/2024, you invest £1,000 on a one-month Gilt with a rate of 5.25%

$$\text{Payoff on } 30/11/2024 = \$1,000 \times \left(1 + \frac{5.25}{100} \times \frac{30}{365}\right) = 1004.315$$

To make our life easier, we will work with 30/360 unless otherwise specified

# Simple Excess Return

The simple excess return between times $t$ and $t+1$ is calculated as

$$R^e_{t+1} = R_{t+1} - R^f_t$$

$R_{t+1}$ and $R^f_t$ are both defined over the same investment horizon

or alternatively as

$$R^e_{t+1} = (R_{t+1} - 1) - (R^f_t - 1)$$

Warning: The excess return is never a gross return, irrespective of using gross or net returns.

# Simple Returns: Dividend Payment

With a dividend payment, the one-period gross simple return is generally calculated as

$$R_{t+1} = \frac{P_{t+1} + D_{t+1}}{P_t} \longrightarrow \text{Dividend Payment}$$

We can also decompose the gross simple return as follows

$$\underbrace{R_{t+1}}_{\text{cum-dividend stock return}} = \frac{D_{t+1}}{P_t} + \underbrace{\frac{P_{t+1}}{P_t}}_{\text{ex-dividend stock return}}$$

In practice, dealing with dividend payments can be complicated as they are paid infrequently. So, we rely on the data provider to make available prices adjusted for dividends.

# Simple Returns: Dividend Payment

**Nikkei 225 (^N225)**
Osaka - Osaka Delayed Price. Currency in JPY

Follow

**33,377.42** +89.12 (+0.27%)
At close: January 5 03:15PM JST

Summary | Chart | Conversations | **Historical Data** | Options | Components



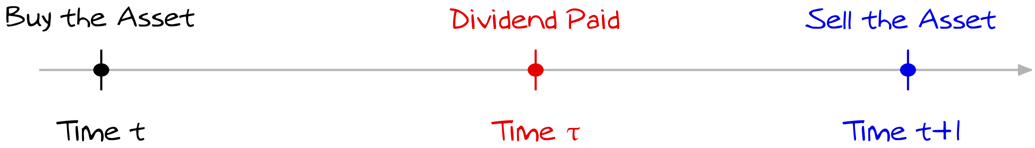
Time Period: Dec 28, 2023 - Jan 08, 2024 Show: Historical Prices Frequency: Daily Apply

Currency in JPY Download

| Date | Open | High | Low | Close* | Adj Close** | Volume |
|---|---|---|---|---|---|---|
| Jan 05, 2024 | 33,397.52 | 33,568.04 | 33,257.43 | 33,377.42 | 33,377.42 | 115,300,000 |
| Jan 04, 2024 | 33,193.05 | 33,299.39 | 32,693.18 | 33,288.29 | 33,288.29 | 117,300,000 |
| Dec 29, 2023 | 33,458.64 | 33,652.71 | 33,305.17 | 33,464.17 | 33,464.17 | 89,000,000 |
| Dec 28, 2023 | 33,477.47 | 33,571.73 | 33,411.24 | 33,539.62 | 33,539.62 | 73,300,000 |

*Close price adjusted for splits. **Adjusted close price adjusted for splits and dividend and/or capital gain distributions.

# Simple Returns: Dividend Payment




Suppose the dividend is paid at time $\tau$ but kept in cash until time $t+1$

cum-dividend stock return

$$\boxed{R_{t+1}} = \frac{D_\tau}{P_t} + \boxed{\frac{P_{t+1}}{P_t}}$$

ex-dividend stock return

Identical to the traditional formula

# Simple Returns: Dividend Payment

The dividend paid at time $\tau$ is immediately reinvested in the same asset

cum-dividend stock return

$$\boxed{R_{t+1}} = \frac{P_t \frac{P_{t+1}}{P_t} + D_\tau \frac{P_{t+1}}{P_\tau}}{P_t}$$

simplify

$$= \left(1 + \frac{D_\tau}{P_\tau}\right) \boxed{\frac{P_{t+1}}{P_t}}$$

ex-dividend stock return

# Simple Returns: Dividend Payment

The dividend paid at time $\tau$ is immediately reinvested in a zero-coupon bond

cum-dividend stock return

$$\boxed{R_{t+1}} = \frac{P_t \frac{P_{t+1}}{P_t} + D_\tau \frac{B_{t+1}}{B_\tau}}{P_t}$$

simplify

$$= \left(1 + \frac{D_\tau}{P_{t+1}} \boxed{\frac{B_{t+1}}{B_\tau}}\right) \boxed{\frac{P_{t+1}}{P_t}}$$

gross bond return
(e.g., deposit rate from $\tau$ to $t+1$)

ex-dividend stock return

# Simple Portfolio Return

Consider $n$ assets and hold them for a single period $(i = 1, 2, \ldots, n)$

- $P_{i,t}$ is the price of the asset $i$,
- $Q_i$ is the number of the asset $i$.

The value of the portfolio at time $t$ is

$$V_t = Q_1 P_{1,t} + Q_2 P_{2,t} + \ldots + Q_n P_{n,t}$$

The portfolio weight of asset $i$ at time $t$ is

$$w_{i,t} = \frac{Q_i P_{i,t}}{V_t}$$

# Simple Portfolio Return

The value of the portfolio at time $t + 1$ is

$$V_{t+1} = Q_1 P_{1,t+1} + Q_2 P_{2,t+1} + \ldots + Q_n P_{n,t+1}$$

Replace

$$Q_i = V_t \frac{w_{i,t}}{P_{i,t}}$$

Rewrite as

$$V_{t+1} = \frac{w_{1,t} V_t P_{1,t+1}}{P_{1,t}} + \frac{w_{2,t} V_t P_{2,t+1}}{P_{2,t}} + \ldots + \frac{w_{n,t} V_t P_{n,t+1}}{P_{n,t}}$$

$$= V_t (w_{1,t} R_{1,t+1} + w_{2,t} R_{2,t+1} + \ldots + w_{n,t} R_{n,t+1})$$

# Simple Portfolio Return

The simple portfolio return at time $t+1$ is given by

$$R^p_{t+1} = \frac{V_{t+1}}{V_t}$$

$$= w_{1,t}R_{1,t+1} + w_{2,t}R_{2,t+1} + \ldots + w_{n,t}R_{n,t+1}$$

Simple returns have the property of portfolio additivity since

$$R^p_{t+1} = \sum_{i=1}^{n} w_{i,t}R_{i,t+1}$$

# Log (or Compounded) Returns

# Log Returns: One Period

With a simple return, gains/losses are accrued over discrete intervals (e.g., every day)

$$R_{t+1} = \frac{P_{t+1}}{P_t}$$

With a log return, gains/losses are accrued continuously (e.g., every millisecond)

I could have used $r_{t,t+1}$ but we omit the first $t$ for simplicity

$$r_{t+1} = \ln(R_{t+1})$$

Why t+1 rather than t? Because the return is known ex-post at t+1

# Log Returns: Two Periods

The log return between times $t$ and $t+2$ is calculated as

$$r_{t,t+2} = \ln\left(\frac{P_{t+2}}{P_t}\right) = \ln\left(\underbrace{\frac{P_{t+2}}{P_{t+1}}}_{\text{one-period gross return } R_{t+2}} \times \underbrace{\frac{P_{t+1}}{P_t}}_{\text{one-period gross return } R_{t+1}}\right)$$

The log return between times $t$ and $t+2$ can be then written as

$$r_{t,t+2} = \ln(R_{t+2}) + \ln(R_{t+1})$$

the sum of one-period gross returns in log terms.

# Log Returns: Multi Periods

The log return between times $t$ and $t+k$ is calculated as

$$r_{t,t+k} = \ln\left(\frac{P_{t+k}}{P_t}\right) = \ln\left(\frac{P_{t+k}}{P_{t+k-1}} \times \frac{P_{t+k-1}}{P_{t+k-2}} \times \ldots \times \frac{P_{t+1}}{P_t}\right)$$

one-period gross return $R_{t+k}$

one-period gross return $R_{t+k-1}$

one-period gross return $R_{t,t+1}$

The log return between times $t$ and $t+k$ can be then written as

$$r_{t,t+k} = \sum_{i=1}^{k} \ln(R_{t+i})$$

the sum of one-period gross returns in log terms.

# Average Log Returns

The average log return between times $t$ and $t+k$ can be then written as

$$\bar{r} = \frac{1}{k}\sum_{i=1}^{k} \ln(R_{t+i})$$

the average of one-period gross returns in log terms being them time additive.

# Log Returns: Other Definitions

The one-period log return between times $t$ and $t+1$ with a dividend payment is

$$r_{t+1} = \ln\left(\frac{P_{t+1} + D_{t+1}}{P_t}\right)$$

$$= \ln(P_{t+1} + D_{t+1}) - \ln(P_t)$$

The one-period log return between times $t$ and $t+1$ in real terms

$$r^r_{t+1} = \ln\left(\frac{P_{t+1}/P_t}{CPI_{t+1}/CPI_t}\right)$$

$$= \ln\left(\frac{P_{t+1}}{P_t}\right) - \ln\left(\frac{CPI_{t+1}}{CPI_t}\right)$$

$$= r_{t+1} - \pi_{t+1}$$

# Log Riskless Return

The log riskless rate between times $t$ and $t+1$ is calculated as

$$r_t^f = \ln B_{t+1} - \ln B_t \longrightarrow$$

Why $t$ rather than $t+1$?
Because the return is
known ex-ante at $t$

Practically, the log riskless return between times $t$ and $t+1$ is calculated as

continuously
compounded
interest rate

$$\longleftarrow r_t^f = \ln\left(1 + \frac{\text{quoted rate}}{100} \times \frac{\text{day count}}{\text{convention}}\right) \longrightarrow \ln R^f$$

# Log Returns: Other Definitions

The one-period log excess return between times $t$ and $t+1$ is

$$r^e_{t+1} = \ln(R_{t+1}) - \ln(R^f_t)$$

Wrong to calculate as
$\ln(R_{t+1} - R^f_t)$

or alternatively as

$$r^e_{t+1} = r_{t+1} - r^f_t$$

This is the log
riskless return

The portfolio log excess return between times $t$ and $t+1$ is calculated as

$$r^p_{t+1} = ln(R^p_{t+1}) = ln\left(\sum_{i=1}^n w_{i,t} R_{i,t+1}\right) \neq \left(\sum_{i=1}^n w_i r_{i,t+1}\right)$$

Log of weighted average
gross simple returns

weighted average
of log returns

# Log Returns: Other Definitions

When returns are small ('measured over short time intervals')

$$r_{t+1}^{p} \approx \left(\sum_{i=1}^{n} w_i r_{i,t+1}\right)$$

When do we use log returns?

# From Simple to Log Returns

Recall the second-order Taylor expansion for $f(x)$ around $a$

$$f(x) \approx f(a) + f'(a)(x-a) + \frac{f''(a)}{2!}(x-a)^2$$

By setting $x = R$, $f(x) = \ln(R)$ and $a = 1$, we have that

$$f(R) = \ln(R) \implies \ln(1) = 0$$

$$f'(R) = \frac{1}{R} \implies f'(1) = 1$$

$$f''(R) = -\frac{1}{R^2} \implies f''(1) = -1$$

# From Simple to Log Returns

Substitute and obtain that

$$\underbrace{r_{t+1}}_{\text{log return}} \approx \underbrace{(R_{t+1}-1)}_{\text{simple return}} - \frac{1}{2}\underbrace{(R_{t+1}-1)^2}_{\text{squared simple return}}$$

Since the variance term is always positive, it follows

log return $\leq$ simple return

# From Simple to Log Returns

Let's take the expectation on both sides

$$\mathbb{E}(r_{t+1}) \approx \mathbb{E}(R_{t+1} - 1) - \frac{1}{2}\mathbb{E}[(R_{t+1} - 1)^2]$$

Recall that

$$\mathbb{V}\mathrm{ar}(R_{t+1} - 1) = \mathbb{E}[(R_{t+1} - 1)^2] - [\mathbb{E}(R_{t+1} - 1)]^2$$

squared average return is small

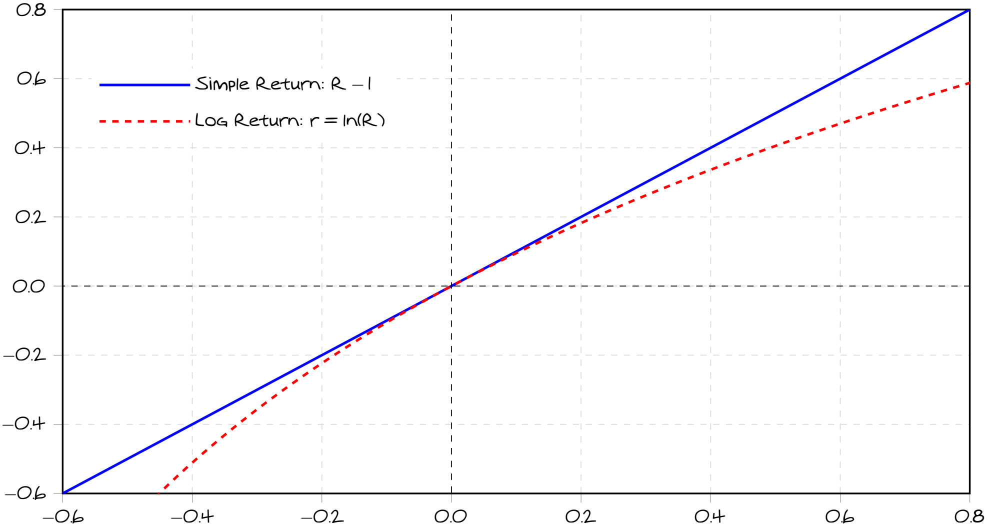
# From Simple to Log Returns

Let's take the expectation on both sides

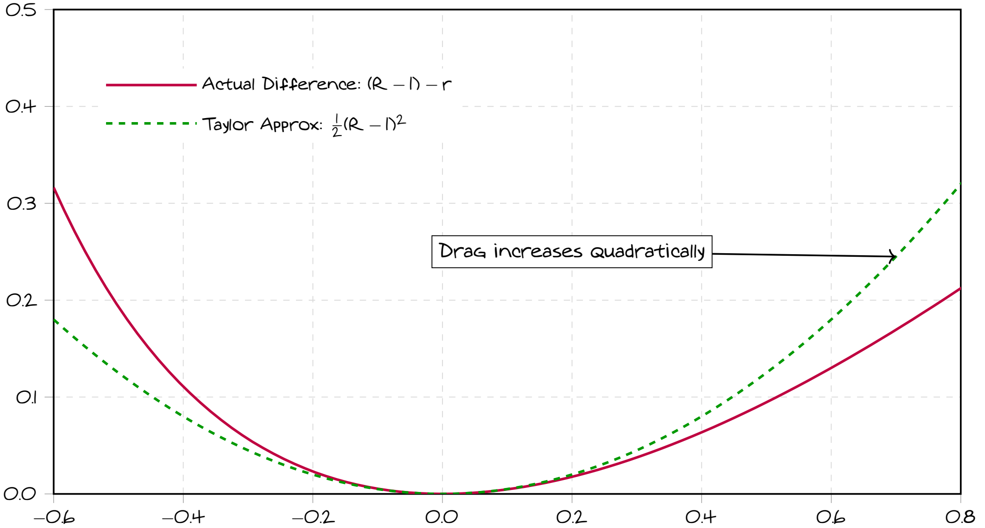
$$\underbrace{\mathbb{E}(r_{t+1})}_{\text{average log return}} \approx \underbrace{\mathbb{E}(R_{t+1}-1)}_{\text{average simple return}} - \frac{1}{2}\underbrace{\mathbb{V}\mathrm{ar}(R_{t+1}-1)}_{\text{variance simple return}}$$

Higher variance in simple returns "drags" down average log returns.

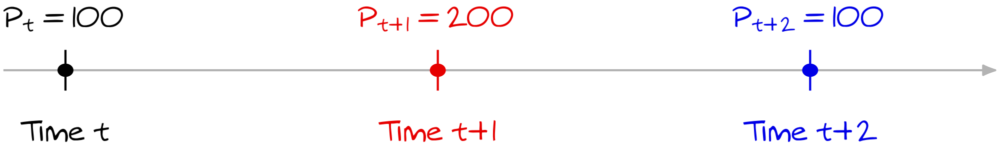
# Simple vs Log Returns



# Volatility Drag



# Volatility Drag: An Example




With simple returns:

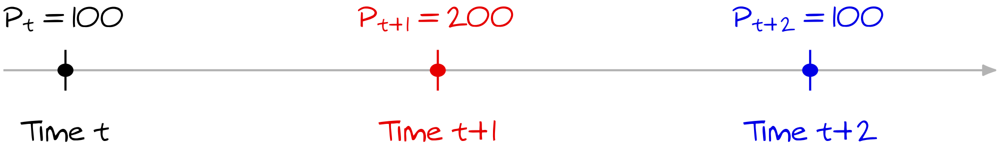
$$R_{t+1} - 1 = 100\%$$

$$R_{t+2} - 1 = -50\%$$

$$Average = \frac{100 - 50}{2} = 25\% \longrightarrow$$

Wrong conclusion: The stock didn't go up in value

# Volatility Drag: An Example




With log returns:

$$r_{t+1} = 69.3\%$$

$$r_{t+2} = -69.3\%$$

$$Average = \frac{69.3 - 69.3}{2} = 0\% \longrightarrow$$

Correct conclusion:
The travel distance
of the stock was none

# Stylized Facts

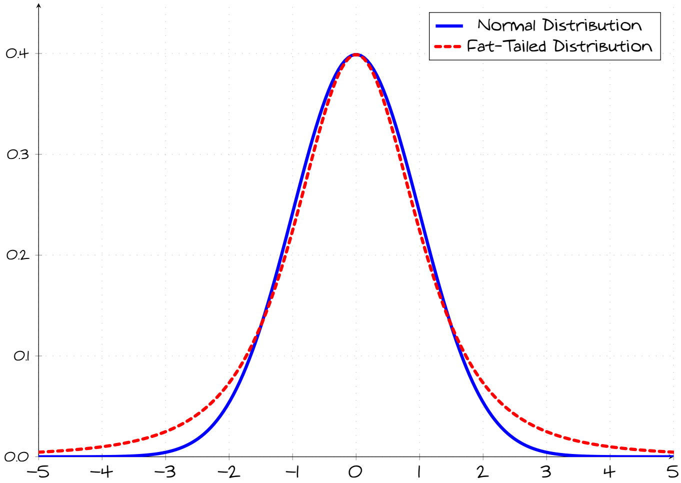
# Stylized Facts

Empirical studies have identified several stylized facts among financial data

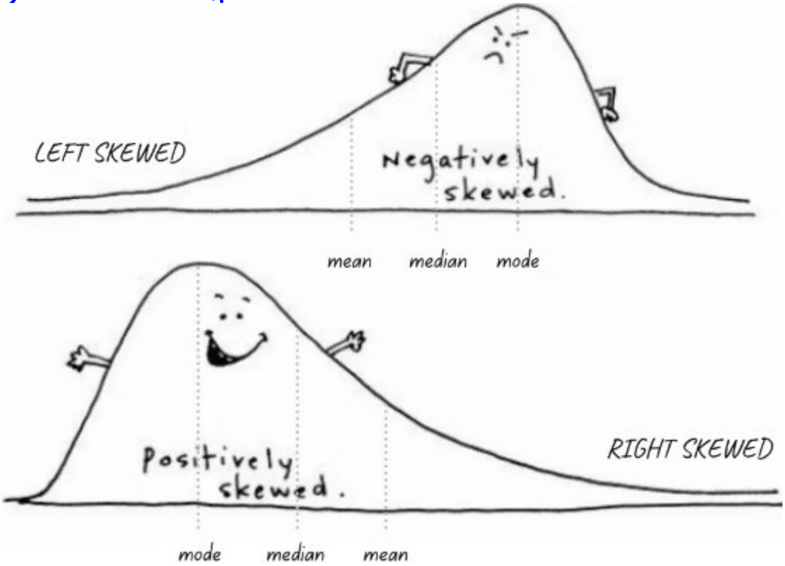
1. Fat Tails $\longrightarrow$ The unconditional distribution of returns displays fatter tails than the normal distribution. Using a normal distribution to model returns, we underestimate both the number and the magnitude of booms and busts.

2. Asymmetry $\longrightarrow$ The unconditional distribution is negatively skewed suggesting that extreme negative returns are more frequent than extreme positive returns. The asymmetry and the fat-tails persists even after adjusting for conditional heteroskedasticity. This means that the conditional distribution is also non-normal.

# Fat Tails: An Example



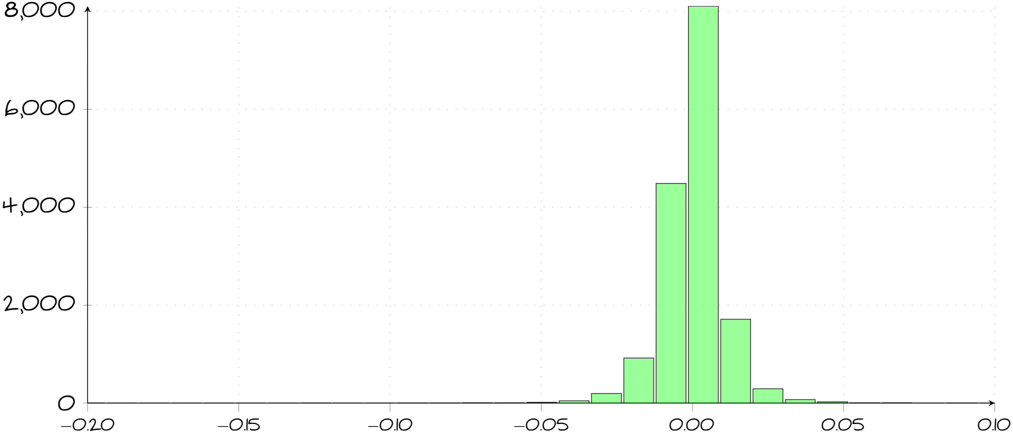
# Asymmetry: An Example



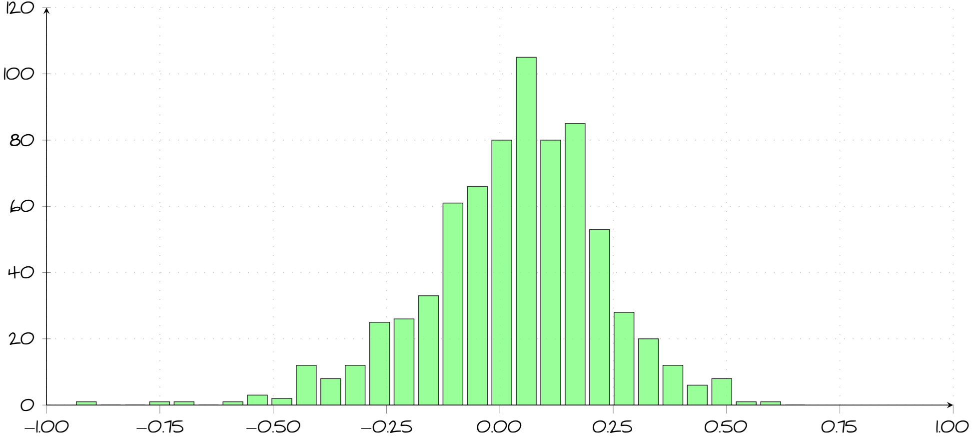
# Stylized Facts

3. Aggregate Normality $\longrightarrow$ As the frequency of the returns diminishes, the distribution of financial returns approaches the normal distribution.
4. Absence of Serial Correlation $\longrightarrow$ Financial returns display little or no serial correlations, except at very high frequency.
5. Volatility Clustering $\longrightarrow$ Volatility of returns is serially correlated, suggesting that a large positive (negative) return tends to be followed by another large positive (negative) return.
6. Time-varying Correlations $\longrightarrow$ Correlations between asset returns tends to increase during high-volatility periods, in particular during crashes.

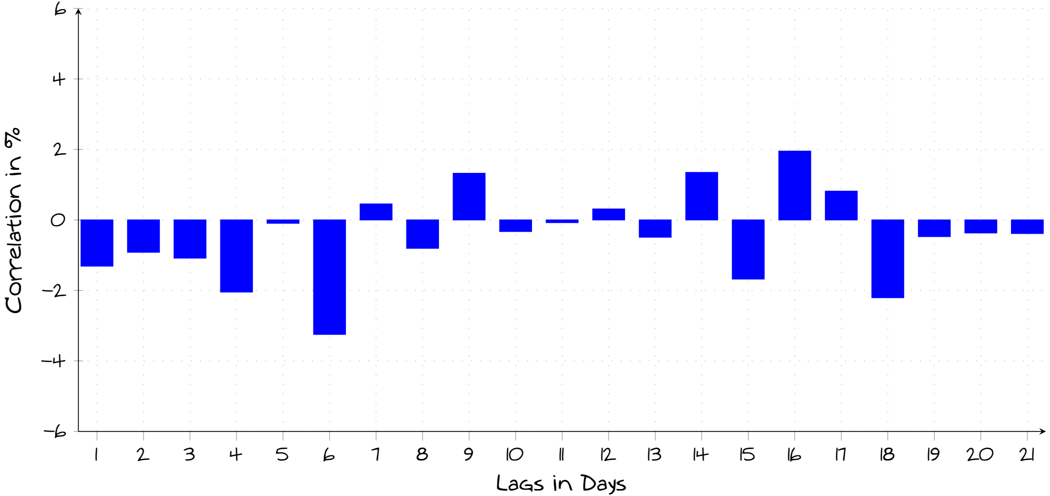
# Aggregate Normality: S&P500 Daily Returns



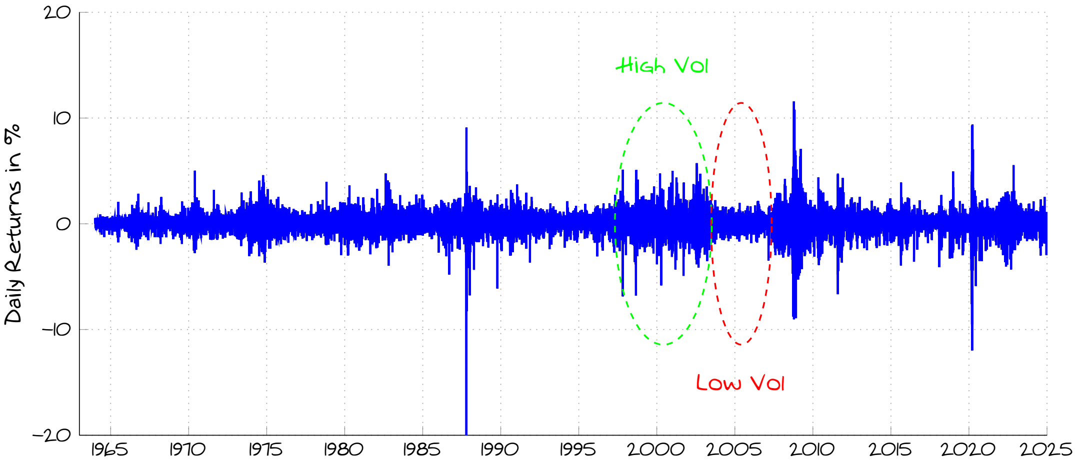
# Aggregate Normality: S&P500 Monthly Returns



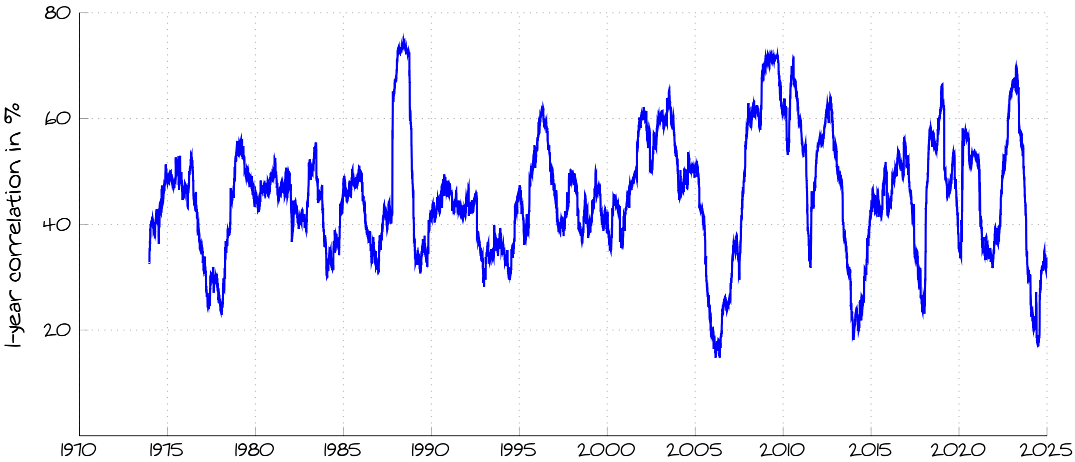
# Serial Correlation: S&P500 Daily Returns



# Clustering: S&P500 Daily Returns



# Time-Varying Correlation: HP and Intel Daily Returns



# Moments

# Probability Density Function

Let $x$ be a continuous random variable and $f(x)$ its probability density function $(pdf)$

- $f(x)$ describes the probability of different values such that for any interval $[a, b]$

$$Pr(A \leq x \leq B) = \int_A^B f(x)dx$$

where $Pr(\cdot)$ is the *area under the probability curve over the interval* $[A, B]$

- $f(x)$ must satisfy the following condition

$$\int_{-\infty}^{\infty} f(x)dx = 1$$

# Moments

The $k^{th}$ uncentered moment of $x$ is defined as

$$m_k = E\left[x^k\right] = \int_{-\infty}^{\infty} x^k f(x)dx$$

where $k$ is an integer.

The $k^{th}$ centered moment of $x$ is defined as

$$\mu_k = E\left[(x - m_1)^k\right] = \int_{-\infty}^{\infty} (x - m_1)^k f(x)dx$$

where $m_1 = E\left[x\right]$ is the first uncentered moment.

# Moments

Uncentered and Centered Moments (if they exist) are related via the binomial transformation

$$\mu_k = \sum_{i=0}^{k} (-1)^{k-i} \binom{k}{i} m_i m_1^{k-i}$$

The binomial term

$$\binom{k}{i} = \frac{k!}{i!(k-i)!}$$

determines how much weight each uncentered moment contributes to the centered moment.

# Moments

To sum up, we have the following

$$\mu_1 = 0$$
$$\mu_2 = m_2 - m_1^2$$
$$\mu_3 = m_3 - 3m_2m_1 + 2m_1^3$$
$$\mu_4 = m_4 - 4m_3m_1 + 6m_2m_1^2 - 3m_1^4$$

where

$\mu_1 \longrightarrow$ zero by construction

$\mu_2 \longrightarrow$ this is the variance

$\mu_3 \longrightarrow$ non-standardized skewness

$\mu_4 \longrightarrow$ non-standardized kurtosis

# Moments

The mean (first non-centered moment) is an indicator of central tendency

$$\mu = E\left[x\right] = m_1$$

The variance (second centered moment) is an indicator of dispersion around $\mu$

$$\sigma^2 = V\left[x\right] = m_2 - m_1^2$$

# Moments

The skewness (standardized third centered moment) is an indicator of symmetry

$$sk = Skewness[x] = E\left[\left(\frac{x-\mu}{\sigma}\right)^3\right]$$

The kurtosis (standardized fourth centered moment) is an indicator of tail heaviness

$$ku = Kurtosis[x] = E\left[\left(\frac{x-\mu}{\sigma}\right)^4\right]$$

# Empirical Moments

How do practically estimate the moments?

- Let $\{x_1, \ldots, x_T\}$ be a random sample of $T$ observations.

The sample mean is

$$\widehat{\mu} = \frac{1}{T} \sum_{i=1}^{T} x_t,$$

The sample variance is

$$\widehat{\sigma}^2 = \frac{1}{T-1} \sum_{i=1}^{T} (x_t - \widehat{\mu})^2$$

# Empirical Moments

The sample skewness is

$$\widehat{s} = \frac{1}{T-1} \sum_{i=1}^{T} \frac{(x_t - \widehat{\mu})^3}{\widehat{\sigma}^3}$$

The sample kurtosis is

$$\widehat{\kappa} = \frac{1}{T-1} \sum_{i=1}^{T} \frac{(x_t - \widehat{\mu})^4}{\widehat{\sigma}^4}$$

# Conventional Moments

### Problems with traditional measures

- Kim & White (2004) show that conventional skewness and kurtosis are not accurate,
- A single extreme observation can drastically distort conventional skewness and kurtosis
- Why? Noisy sample means and cubing/fourth powers amplifying outlier effects.

### Kim & White (2004) propose robust alternatives

- Let $Sk_1$ be the traditional measure of skewness,
- Let $Ku_1$ be the traditional measure of kurtosis.

# Robust Measures of Skewness

## Bowley's coefficient of skewness

Measures the imbalance between upper and lower parts of the distribution

$$Sk_2 = \frac{Q_3 + Q_1 - 2Q_2}{Q_3 - Q_1}$$

Interquartile range acting as a scale factor making it independent of the data's scale

- $Q_1 \longrightarrow$ first quartile,
- $Q_2 \longrightarrow$ median,
- $Q_3 \longrightarrow$ third quartile.

## Interpretation

- $Sk_2 = 0 \longrightarrow$ symmetric distributions,
- $Sk_2 > 0 \longrightarrow$ when upper tail stretches more than lower tail,
- $Sk_2 < 0 \longrightarrow$ when lower tail stretches more than upper tail.

# Robust Measures of Skewness

Groeneveld & Meeden's coefficient of skewness

Measures the distance between mean and median of the distribution

$$Sk_3 = \frac{\mu - Q_2}{E|x - Q_2|}$$

Average absolute deviation from the median. Bounded between -1 and 1

- $\mu \rightarrow$ mean,
- $Q_2 \rightarrow$ median.

Interpretation

- $Sk_3 = 0 \longrightarrow$ symmetric distributions,
- $Sk_3 > 0 \longrightarrow$ positive skewness,
- $Sk_3 < 0 \longrightarrow$ negative skewness.

# Robust Measures of Skewness

Pearson's coefficient of skewness

$$Sk_4 = \frac{\mu - Q_2}{\sigma}$$

Measures the distance between mean and median of the distribution

Normalized by the standard deviation

- $\mu \to$ mean,
- $Q_2 \to$ median,
- $\sigma \to$ standard deviation.

Interpretation

- $Sk_4 = 0 \longrightarrow$ symmetric distributions,
- $Sk_4 > 0 \longrightarrow$ positive skewness,
- $Sk_4 < 0 \longrightarrow$ negative skewness.

# Robust Measures of Kurtosis

Measures the spread in the tails

## Moors' coefficient of kurtosis

$$Ku_2 = \frac{(E_7 - E_5) + (E_3 - E_1)}{E_6 - E_2} - 1.23$$

Centered to the N(0,1)

Measures the spread in the center

- $E_k \longrightarrow k$-th octile of the distribution (octiles divide data into eight equal parts),
- $E_1 \longrightarrow 1^{st}$ octile (12.5%), $E_2 \longrightarrow 2^{nd}$ octile (25%), ...., $E_7 \longrightarrow$,
- $E_6 - E_2 \longrightarrow$ Interquartile range.

## Interpretation

- $Ku_2 = 0 \longrightarrow$ kurtosis similar to the standard normal distribution,
- $Ku_2 > 0 \longrightarrow$ heavy tails,
- $Ku_2 < 0 \longrightarrow$ light tails.

# Robust Measures of Kurtosis

A constant to center the coefficient for the N(0,1)

## Hogg's coefficient of kurtosis

$$Ku_3 = \frac{U_{0.05} - L_{0.05}}{U_{0.5} - L_{0.5}} - 2.59$$

- $U_{0.05} \longrightarrow$ Mean of the largest 5% of the data,
- $L_{0.05} \longrightarrow$ Mean of the smallest 5% of the data,
- $U_{0.5} \longrightarrow$ Median of the upper half of the data,
- $L_{0.5} \longrightarrow$ Median of the lower half of the data.

## Interpretation

- $Ku_3 = 0 \longrightarrow$ kurtosis similar to the standard normal distribution,
- $Ku_3 > 0 \longrightarrow$ heavy tails,
- $Ku_3 < 0 \longrightarrow$ light tails.

# Robust Measures of Kurtosis

A constant to center the coefficient for the N(0,1)

## Crow & Siddiqui's coefficient of kurtosis

$$Ku_4 = \frac{F^{-1}(0.975) + F^{-1}(0.025)}{F^{-1}(0.75) - F^{-1}(0.25)} - 2.91$$

- $F^{-1}(\alpha) \longrightarrow$ quantile of the data.

## Interpretation

- $Ku_4 = 0 \longrightarrow$ kurtosis similar to the standard normal distribution,
- $Ku_4 > 0 \longrightarrow$ heavy tails,
- $Ku_4 < 0 \longrightarrow$ light tails.

# Robust Measures of Skewness and Kurtosis

Kim & White (2004) show that

- $Sk_1$ and $Ku_1$ are severely biased even by a single outlier,
- $Sk_2$ and $Ku_2$ are not distorted by extreme outliers,
- $Sk_3$ and $Ku_3$ take into account the outliers while removing the distortion,
- $Sk_4$ and $Ku_4$ take into account the outliers while removing the distortion.
- $Ku_3$ has a finite sample bias that disappears with $N > 500$.

# Estimates Robust to Outliers

# OLS Estimator

Consider the traditional linear model with a single regressor

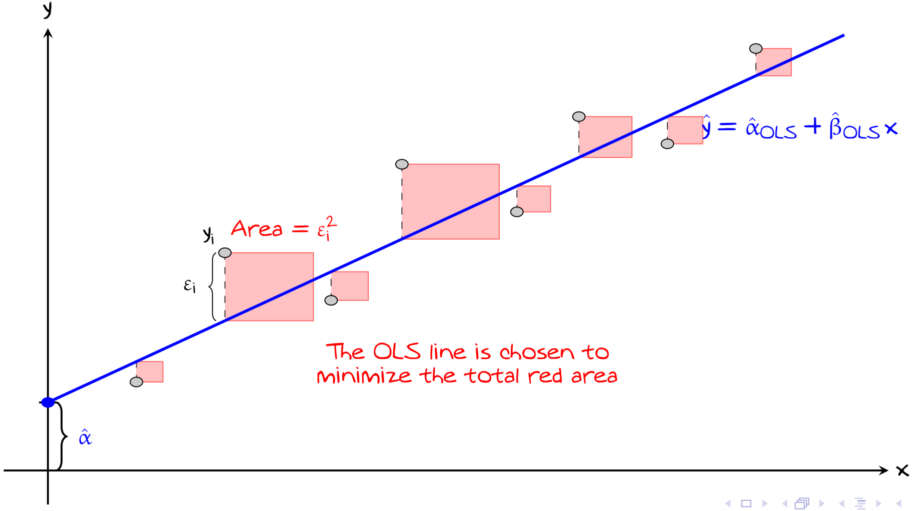
$$\underbrace{y_i}_{\text{dependent variable}} = \underbrace{\alpha}_{\text{coefficients}} + \underbrace{\beta}_{\text{coefficients}} \underbrace{x_i}_{\text{regressor}} + \underbrace{\varepsilon_i}_{\text{unobserved errors}}$$

OLS estimates $\beta$ by minimizing the sum of squared residuals as

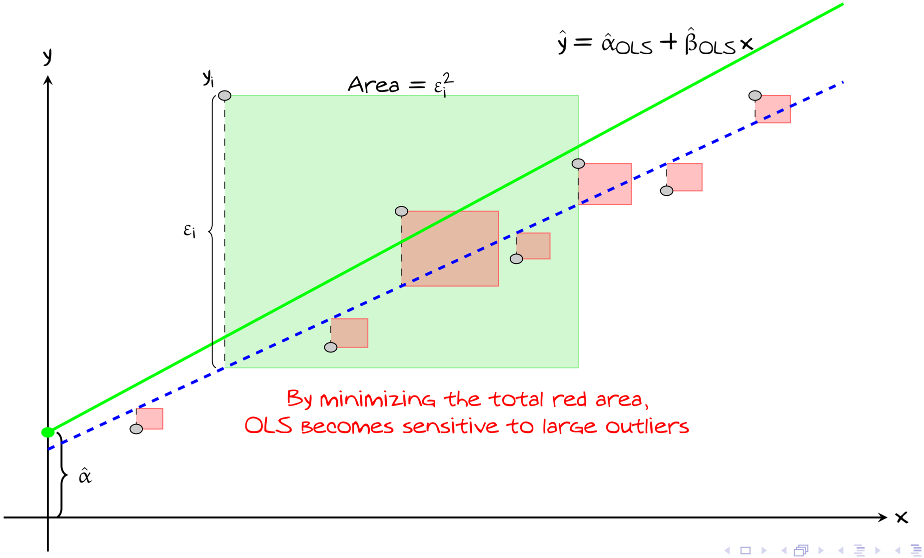
$$\hat{\beta}_{\text{OLS}} = \arg\min_{\beta} \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2$$

Larger errors receive more weight, so a few outliers can dominate the fit!

# OLS Estimator: Ideal Case



# OLS Estimator: Distorted with Outliers



# LAD Estimator

Financial data often exhibits Fat tails, jumps, structural breaks, and data errors

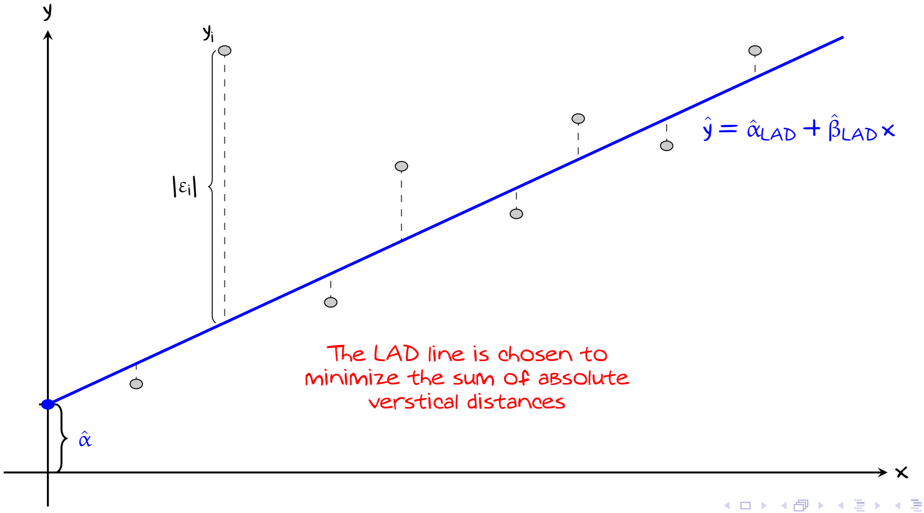
- OLS estimates can be distorted with large outliers.
- A solution is to use the Least Absolute Deviation (LAD) estimator.

LAD estimates $\beta$ by minimizing the sum of absolute residuals as

$$\hat{\beta}_{\text{LAD}} = \arg\min_{\beta} \sum_{i=1}^{n} |y_i - \alpha - \beta x_i|$$

Errors are penalized linearly and no single observation can dominate the fit!

# LAD Estimator: Robust to Outliers



# The Statistical Trade-Off

| Feature | OLS Estimator | LAD Estimator |
|---|---|---|
| Robustness | Low: A single outlier can pull the fitted line substantially because of squared errors. | High: The line effectively ignores outliers, passing through the majority of data. |
| Efficiency | High: If errors are normal, OLS is the most precise estimator possible (BLUE). | Low: Requires larger samples to match OLS precision when data are clean. |
| Computation | Fast and exact: Closed-form solution. | Slow and iterative: No closed form. |

Use OLS as default for 'clean' data but switch to LAD when you detect significant outliers or heavy tails.

# A Simulated Example

```
# Set the seed for reproducibility
Random.seed!(1234)

# Settings
nobs        = 50;                    # Number of observations
true_alpha = 1.0;                    # True alpha
true_beta  = 2.0;                    # True beta

# 2. Simulate Data: Baseine Scenario
x_base = zeros(nobs)
for t in 2:nobs x_base[t] = 0.90* x_base[t-1] + rand(Normal(0, 1)) end
e_base      = rand(Normal(0,  1), nobs);
y_base      = true_alpha .+ true_beta .* x_base + e_base;

# 3. Create a DataFrame and estimate via OLS
df_base     = DataFrame(x = x_base, y = y_base);
ols_base    = lm(@formula(y ~ x), df_base);


```

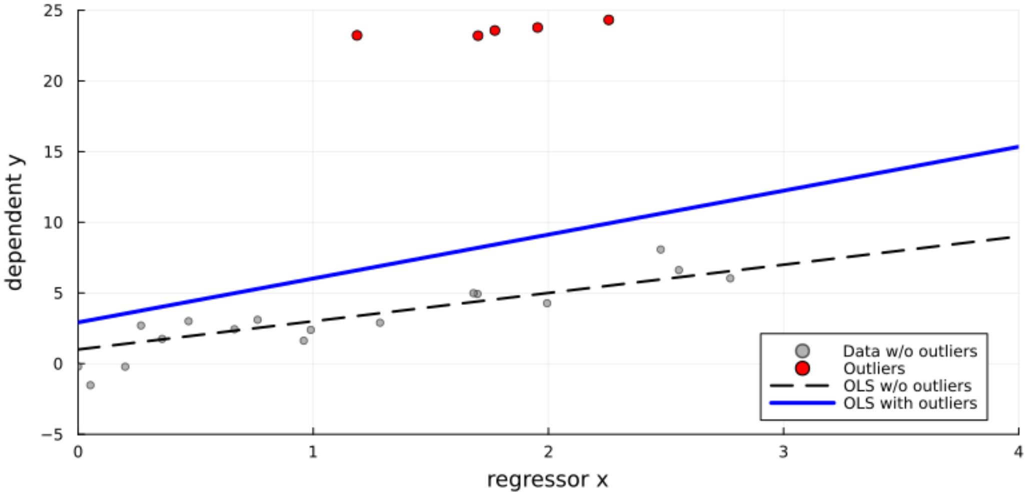
# A Simulated Example

```
# 4. Outliers Scenario
y_outs      = copy(y_base)
y_outs[11] = y_outs[11] + 20;
y_outs[12] = y_outs[12] + 20;
y_outs[13] = y_outs[13] + 20;
y_outs[14] = y_outs[14] + 20;
y_outs[15] = y_outs[15] + 20;


# Update DataFrame and estimate via OLS
df_outs      = DataFrame(x = x_base, y = y_outs);
ols_outs     = lm(@formula(y ~ x), df_outs);


```

# OLS Estimator: Distorted with Outliers



# A Simulated Example

```
# 4. Outliers Scenario
y_outs       = copy(y_base)
y_outs[11] = y_outs[11] + 20;
y_outs[12] = y_outs[12] + 20;
y_outs[13] = y_outs[13] + 20;
y_outs[14] = y_outs[14] + 20;
y_outs[15] = y_outs[15] + 20;


# Update DataFrame and estimate via OLS
df_outs      = DataFrame(x = x_base, y = y_outs);
lad_outs     = qreg(@formula(y ~ x), df_outs, 0.5);


```

# LAD Estimator: Robust to Outliers