TIØ4317: Empirical and Quantitative Methods in Finance

Jacob Clements

January 2025

1 Things to add

- Non-stationarity can lead to spurios regressions.
- IC consists of two components, a goodness of fit statistic and a penalty term that favours parsimonious specifications. IC's differ in their penalty terms.
- Check for non-linearity use correlellogram of squared series.
- If dataset very large use bic instead of aic

2 Basics of Statistical Inference: Should be very comfortable with this theory

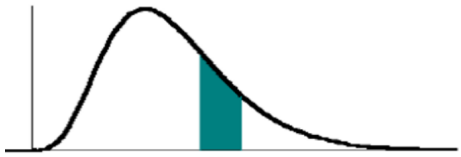
2.1 Random Variables and Distributions:

Every random variable is characterized (fully determined) by its distribution function. We use two types of functions for explaining the random variable.

2.1.1 Probability Density Function (PDF)

$$Pr(a \leq X \leq b) = \int_a^b f(x)dx$$

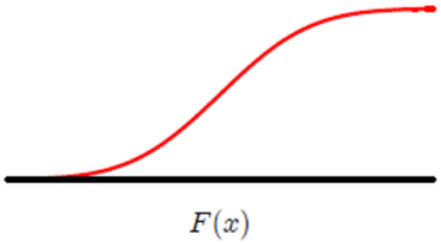
- It is defined so that $f(x) \geq 0$
- The total area under the PDF is 1
- The probability that the random variable takes values between a and b is the higlighted are under the pdf:



2.1.2 Cumulative Distribution Function (CDF)

$$Pr(X \leq x) = F(x)$$

- $F(x)$ is a non-decreasing function
- $lim_{x \to -\infty} F(x) = 0$
- $lim_{x \to \infty} F(x) = 1$
- For a continuous random variable, the CDF is the integral of the PDF.



2.2 Moments of a Distribution:

- "Moment" of a distribution refers to how probability mass is distributed.
- The **raw kth moment** of a function $f(x)$ is a moment defined as

$$E[X^k]$$

- **Generalization:** the kth moment of a function $f(x)$ about a non-random value c is

$$E[(X - c)^k] = \int_{-\infty}^{\infty} (x - c)^k f(x)\, dx$$

- Distinction here: a **raw moment** is a moment about the origin ($c = 0$), and a **central moment** is a moment about the distribution's mean ($c = E[X]$).
  - Useful because they allow us to quantify properties of distributions in ways that are **location-invariant**.

- Finally, the k**th standardized moment** is defined as the kth central moment normalized by the standard deviation. If $E[X] = \mu$ and σ is the standard deviation of X:

$$E\left[\left(\frac{X-\mu}{\sigma}\right)^k\right]$$

- **Note!** Only well-defined for distributions whose first two moments exist and second moment is non-zero. Standardization makes the moment both location- and scale-invariant.

2.2.1 Usefulness of Moments

Different moments give different insights about the distribution:

- **First Moment** ($k = 1$): The mean ($E[X]$), showing where the distribution is centered.
- **Second Central Moment** ($k = 2$): The variance ($E[(X - E[X])^2]$), measuring how spread out the values are around the mean.
- **Third Central Moment** ($k = 3$): Skewness, indicating whether the distribution leans to the left or right of the mean.
- **Fourth Central Moment** ($k = 4$): Kurtosis, describing the "tailedness" of the distribution (how extreme the outliers are).

Examples:

- **Example 1:** I want to forecast the average return on the S&P500 tomorrow, so I can use my data more efficiently if I just model the average return.
- **Example 2:** I want to forecast the volatility of the S&P500 tomorrow, so I can use my data more efficiently if I just model the variance.
- **Example 3:** I am a risk manager and I want to assess the risk on a portfolio, so I can use my data more efficiently if I model the kurtosis of the portfolio returns.

2.3 Association Between Random Variables:

2.3.1 Covariance

Covariance tells us something about co-varying of random variables: The covariance matrix is the the second order moment about the mean.
For the random varianbles X_1, ..., X_n it is an $n \times n$ matrix with elements:

$$cov(X_i, X_j) = \sigma_{ij} = E((X_i - E[X_i])(X_j - E[X_j]))$$

for $i, j = 1, ..., n$.
σ_{ij} measures the degree of linear association between variables:

- Diagonal of covariance matrix contains the variances $\sigma_{ii} = \sigma_i^2$
- Off-diagonal of covariances contains the covariances σij
- Covariance matrix is symmetric $\sigma_{ij} = \sigma_{ji}$
- Covariance matrix is positive semi-definite

2.3.2 Correlation

A unit-free measure of linear association is the correlation coefficient, which is often denoted by ρ and is defined as:

$$\rho = \frac{Cov(X_i, X_j)}{\sqrt{Var(X_i)Var(X_j)}} = \frac{\sigma_{ij}}{\sigma_i \sigma_j}$$

The diagonal of the correlation matrix contains 1. A variable is always perfectly correlated with itself. The correlation matrix is symmetric.

2.4 Relation Between Probability and Statistics

- In probability we ask:
  - given a data generating process, what are the properties of the outcomes?
- In statistics the question is the reverse:
  - given the outcomes, what can we say about the process that generated the data?
- Statistical inference consists of
  - Estimation (point, interval)
  - Inference (quantifying sampling error, hypothesis testing)

2.5 Statistics

- For the mean of a population, we often use the sample mean as a "good guess".
  - The sample mean is a quantity that can be computed after observing the data. This quantity is called a **statistic**.
  - Before you look at the data, a statistic is a random variable. **Why?** Depending on how we sample, the computed statistic could vary.
  - Any observed value after plugging in the data is a realization.
- So it is important to distinguish between
  - the statistic (a random variable)

  - the realization of the statistic for a given sample (a number)

- We therefore denote the statistic with capitals, and the realization of the statistic with small letters, for example for the sample mean:

$$\bar{X} = \frac{1}{n}\sum_{n} X_i \quad \text{and} \quad \bar{x} = \frac{1}{n}\sum_{n} x_i$$

- In statistical inference we think of data as a realization of a random process.

2.6 The Sampling Distribution

- Statistics vary from sample to sample. The sampling distribution of a statistic is the nature of this variability, it can sometimes be determined and often approximated.
- **Sample distribution:** The distribution of the values we get when computing a statistic in (infinitely) many random samples.
  - More generally the sampling distribution of a statistic depends on the sample size.
- The sampling distribution of the data used to construct the statistic can be complicated!
- We can sometimes learn about the sampling distribution of a statistic by
  - Deriving the finite sample distribution
  - Approximation with a Normal distribution in large samples
  - Approximation through numerical simulation

2.7 Limit Theorems for the Sample Mean

2.7.1 Theorem: Law of Large Numbers (LLN)

Suppose that X_1,...,X_n are independent and identically distributed (iid) random variables with $E(X_i) = \mu$ and $Var(X_i) = \sigma^2$, for all $i = 1, ..., n$. Then,

$$\bar{X} \xrightarrow{p} \mu, \quad \text{as } n \to \infty.$$

2.7.2 Theorem: Central Limit Theorem (CLT)

Suppose that you have independent random variables X_1,...,X_n, all with the same distribution, with mean μ and variance σ. For large n it holds that $\bar{X}$ approximately (or asymptotically) follows a normal distribution. We write,

$$\bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right).$$

2.8 Sampling Distribution of the Sample Mean

- If $X_1, \ldots, X_n$ is a random sample from a population with mean μ and variance σ^2; i.e., $X_1, \ldots, X_n \sim$ i.i.d. (μ, σ^2)
  - Then, we can show that $\bar{X}$ is a random variable with mean μ and variance $\frac{\sigma^2}{n}$ (cannot say anything more than this!).
- Suppose now that $X_1, \ldots, X_n \sim$ i.i.d. $\mathcal{N}(\mu, \sigma^2)$
  - Then $\bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$.
  - Here we can give the full sampling distribution, because a linear combination of normal random variables is normal.
- As $n \to \infty$, we can work with limiting distributions and use asymptotic results (treating them as approximations), here LLN and CLT come in handy.
  - If the statistic is (a function of) a sample average and the sample is large, then we can often approximate the sampling distribution with a Normal distribution.

2.8.1 Proof of Mean and Variance Expressions

$$* \quad E(\bar{X}) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n}\sum_{i=1}^{n} \mu = \mu$$

$$* \quad \text{Var}(\bar{X}) = E\left[\left(\bar{X} - E\bar{X}\right)^2\right] = \frac{1}{n^2} E\left[\left(\sum_{i=1}^{n}(X_i - \mu)\right)^2\right]$$

$$* \quad = \frac{1}{n^2}\left[\sum_{i=1}^{n} E\left((X_i - \mu)^2\right) + \sum_{i \neq j} E\left((X_i - \mu)(X_j - \mu)\right)\right]$$

$$* \quad = \frac{1}{n^2}\sum_{i=1}^{n} \text{Var}(X_i) + 0 \quad \text{by Independence}$$

$$* \quad = \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2 = \frac{\sigma^2}{n}.$$

2.9 Bootstrapping

- We can also use simulations to approximate sampling distributions without a model.
- Such a procedure is called **bootstrap**:
  1. draw a random sample of size n with replacement from our data,
  2. compute our statistic,
  3. repeat steps 1 and 2 many times.
- The resulting distribution of statistics across our resamples is an approximation of the sampling distribution of our statistic.
- The idea is that a random sample of a random sample from the population is, again, a random sample of the population.

2.10 Estimation

- We start as usual with an assumption about how our data came about (the "data generating process").
- We consider observations in our data $X_1, \ldots, X_n$ as draws from a population or, more generally, an unknown probability distribution $f(X)$.
- The aim is to use our sample data $X_1, \ldots, X_n$ (*remember we assume them to be i.i.d.*) to learn something about the unknown probability distribution $f(X)$ where the data came from.
  - We want to estimate a population parameter using the observed data, some measure of variation, an average, min, max, quantile, etc.
- *Point estimation* attempts to obtain a best guess for the value of that parameter.
- An *estimator* is a statistic (a function of data) that produces such a guess.
- A bit more formally: If we are interested in a parameter θ and $X_1, \ldots, X_n$ is a random sample from X. An estimator for a parameter θ is a statistic $\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$.
  - A notational convention is that an estimator for a parameter is denoted with a "hat".
- There are several well-established methods of obtaining estimators and we will go through 2 of the general methods in this subsection.

2.10.1 Maximum Likelihood Estimators

- The basic idea is as follows: we have a statistical model with some unknown parameter θ and "nature" has given us observations drawn from a distribution that depends on θ.
- Our best estimate for θ, then, is to find the value $\hat{\theta}$ that makes *what we have observed the most likely outcome*. This way of estimating uses the idea of likelihood.

$$L(\theta) = L(\theta|x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x_i|\theta)$$

- The likelihood function is nothing more than the joint density (or mass) function of a random sample.
- The maximum likelihood approach prescribes to find the value of θ that maximizes the likelihood function.

$$\hat{\theta}_{ML} = \arg\max_{\theta} L(\theta)$$

- However, it is more common to work with the *log-likelihood*, which is the (natural) logarithm of the likelihood function:

$$\ell(\theta) := \log(L(\theta))$$

- The log-likelihood turns the product into a sum, which makes the optimization easier to solve.

We now show a common example of how this method is used to find estimators:

- Let $(X_1, \ldots, X_n) \sim \text{iid}\, \mathcal{N}(\mu, \sigma^2)$, where both μ and σ^2 are unknown. Find the ML estimators for μ and σ^2.

$$L(\mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x_i-\mu)^2} = (2\pi\sigma^2)^{-\frac{n}{2}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2}$$

$$\ell(\mu, \sigma^2) = \log\left((2\pi\sigma^2)^{-\frac{n}{2}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2}\right)$$

$$= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2$$

$$\frac{\partial \ell(\cdot)}{\partial \mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu) = 0 \quad \iff \quad n\mu = \sum_{i=1}^{n} x_i \quad \iff \quad \mu = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{X},$$

$$\frac{\partial \ell(\cdot)}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i-\mu)^2 = 0 \quad \iff \quad \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2.$$

2.10.2 Method of Moments

- A second general method to find estimators in statistical models is the method of moments.
- The idea: base the estimate of a *population* moment by computing the analogous *sample* moment.
- Even though the basic idea is quite simple, writing down a general definition is quite tricky.
  - This is because we can, in principle, write down as many moments of the distribution of some random variable X as we want (assuming that they exist).
  - In general, these moments will depend on (some of the) unknown parameter(s). So, if we have k unknown parameters $\theta = (\theta_1, \ldots, \theta_k)$, let's denote the m-th moment by $\mu_m(\theta) := E(X^m)$.
  - Then the corresponding sample moment is

$$M_m(X_1, \ldots, X_n) := \frac{1}{n} \sum_{i=1}^{n} X_i^m.$$

- A *method of moments (MM) estimator* for θ is obtained by simultaneously solving k equations:

$$\mu_{m_1}(\hat{\theta}) = M_{m_1}(X_1, \ldots, X_n),$$

$$\vdots$$

$$\mu_{m_k}(\hat{\theta}) = M_{m_k}(X_1, \ldots, X_n).$$

- where all m_i, $1, \ldots, k$ are distinct. Usually, we take the first k moments, but that does not always work.

We now show a common example of how this method is used to find estimators:

Let $(X_i)_{i=1}^n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$. Find the moment estimators for μ and σ^2.
We have that

$$E(X) = \mu := \mu_1(\mu, \sigma), \quad M_1(X_1, \ldots, X_n) = \bar{X},$$

and

$$E(X^2) = \text{Var}(X) + E(X)^2 = \sigma^2 + \mu^2 := \mu_2(\mu, \sigma), \quad M_2(X_1, \ldots, X_n) = \frac{1}{n} \sum_{i=1}^{n} X_i^2.$$

The moment equations are:

$$\begin{cases} \hat{\mu}_{\text{MM}} = \bar{X}, \\ \hat{\sigma}^2_{\text{MM}} + \hat{\mu}^2_{\text{MM}} = \frac{1}{n}\sum_{i=1}^{n} X_i^2. \end{cases}$$

Solving for $\hat{\sigma}^2_{\text{MM}}$, we have:

$$\begin{cases} \hat{\mu}_{\text{MM}} = \bar{X}, \\ \hat{\sigma}^2_{\text{MM}} = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2. \end{cases}$$

Alternatively:

$$\begin{cases} \hat{\mu}_{\text{MM}} = \bar{X}, \\ \hat{\sigma}^2_{\text{MM}} = \frac{1}{n}\sum_{i=1}^{n} (X_i - \bar{X})^2. \end{cases}$$

2.11 Point Estimation

- We usually mean by "best" an estimator whose sampling distribution is more concentrated about the population parameter value compared to other estimators.
  - The relevant issue is NOT the closeness of our estimates to the unknown parameters for a particular realized sample BUT the closeness of our estimates on average over repeated samples.
- Hence, the choice of a specific statistic as an estimator depends on the probabilistic characteristics of this statistic in the context of the sampling distribution.
- We will go over three desirable properties of estimators:
  - Finite sampling properties
    - Unbiasedness
    - Efficiency
  - Large sampling property
    - Consistency

2.11.1 Bias

- One desirable property is that an estimator is, on average, correct, i.e., unbiased:

$$E[\hat{\theta}] = \theta$$

- The *estimand* (θ) must be equal to the mean estimator across many random samples.
- It is important to note that bias is not the difference between the estimate (a number from one sample) and the estimand (this is the estimation error).

- Bias is the average estimation error across (infinitely) many random samples!
- Let $X_1, \ldots, X_n$ be our random sample of a random variable X with expectation μ and variance σ^2. Intuitively, you would estimate μ by the sample mean, $\bar{X}$, and σ^2 by the sample variance, denoted by σ^2_X.
  - However, we can show that $\bar{X}$ is an unbiased estimator for μ, but that σ^2_X is not unbiased for σ^2.

Proving unbiased and biased estimators:

Unbiased sample mean:

$$E(\bar{X}) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n} n\mu = \mu,$$

hence this is an unbiased estimator.

Biased sample variance:

$$E\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2\right] = \frac{1}{n}\sum_{i=1}^{n} E\left[(X_i - \bar{X}_n)^2\right]$$

$$= \frac{1}{n}\sum_{i=1}^{n} E\left[X_i^2 - 2X_i\bar{X}_n + \bar{X}_n^2\right]$$

$$= \frac{n-1}{n}\left(E[X_i^2] - E[X_i]^2\right) = \frac{n-1}{n}\mathrm{Var}(X_i),$$

2.11.2 Efficiency

- Unbiasedness is a desirable property but is rarely used by itself as an estimation criterion.
  - Because there are many unbiased estimators that are poor uses of the data.
- In judging two unbiased estimators, choose the estimator with lower variance.
  - The smaller the variance of an estimator, the smaller the spread of its realizations (its estimates) around its mean (which for an unbiased estimator is the true parameter value). So, an unbiased estimator with a smaller variance is "better".
- **Efficiency:** An estimator is said to be *efficient* if it achieves the smallest possible variance among all unbiased estimators of a parameter.

  - Efficiency compares how well an estimator uses the available data to estimate the parameter.

- In practice, we may need to consider both bias and variance to evaluate the overall performance of an estimator.

- **Motivation:** Restricting attention to unbiased estimators only may lead us to discard estimators that, even though they are biased, might still be useful.

2.11.3 Bias-Variance Trade-Off: MSE

- Suppose we have:
  - An unbiased estimator with a large sampling variance.
  - A biased estimator with a small sampling variance.

- Should we choose our "best" estimator based on bias or variance alone?

- A criterion that recognizes the trade-off between bias and variance is *Mean Squared Error* (MSE):

$$\text{MSE}(\hat{\theta}) = E[(\hat{\theta} - \theta)^2]$$

$$\text{MSE}(\hat{\theta}) = \text{Bias}^2(\hat{\theta}) + \text{Variance}(\hat{\theta}),$$

- **Bias**$(\hat{\theta}) = E[\hat{\theta}] - \theta$: distance between the expectation of the estimator and the parameter.

- **Variance**$(\hat{\theta})$: The variability of the estimator across different samples.

- **Example:**
  - Unbiased Estimator $(\hat{\theta}_1)$: Bias $= 0$, Variance $= 5$. $\text{MSE}(\hat{\theta}_1) = 0^2 + 5 = 5$.
  - Biased Estimator $(\hat{\theta}_2)$: Bias $= 1$, Variance $= 1$. $\text{MSE}(\hat{\theta}_2) = 1^2 + 1 = 2$.

- Efficiency evaluates how well an estimator minimizes variance among unbiased estimators. The MSE, however, balances bias and variance, which can be more practical in many cases. We note that the MSE measures the spread of the estimator around the true parameter value.

2.11.4 Consistency

- Maybe in finite samples we do **NOT** know:
  - whether an estimator is unbiased; or
  - what the finite sampling distribution of our estimator is (need to compute moments / requires existence of moments).

- In this case we work with approximations for large sample sizes.
- Consistent estimators converge to the truth as sample sizes grow large:
  - Or in statistical terms, if the sample gets infinitely large, the probability that the estimator deviates from the true parameter gets infinitesimally small.
  - Notation:
$$\lim_{n\to\infty} \hat{\theta} = \theta \quad (\text{or } \hat{\theta} \xrightarrow{p} \theta).$$

2.12 Interval Estimation

- While an estimator may be unbiased or consistent, a given estimate will never equal the true value.
  - This point estimate does not give a sense of "closeness," while the variance estimate does not give a sense of location.
- We could try to combine both in statements like:

  "We are confident that __ lies somewhere between __ and __"

  where we would like to:
  - give a specific interval, and
  - be precise about *how* confident we are.
- This is the aim of a **confidence interval (CI)**, which is a particular type of probability interval.

2.12.1 Confidence Interval (CI)

- Formally, we define a confidence interval as follows:
$$P\left(\hat{\theta} \in (\ell_\theta, r_\theta)\right) = 1 - \alpha$$
  where we construct estimates of a lower bound ℓ_θ and an upper bound r_θ such that the interval covers the parameter of interest with probability $1 - \alpha$.
- We call $1 - \alpha$ the confidence level.
- CI's are random intervals because they differ across random samples.
- The confidence level is thus a probability relative to the sampling distribution!

- For a random sample of size n, the probability that our estimator is no further than ϵ from θ equals:

$$Pr(\theta - \epsilon < \hat{\theta} < \theta + \epsilon)$$

which is the probability that the random variable $\hat{\theta}$ is in a fixed interval with unknown boundaries. So, we write:

$$Pr(\hat{\theta} - \epsilon < \theta < \hat{\theta} + \epsilon).$$

- Now this is the probability that a random interval $(\hat{\theta} - \epsilon, \hat{\theta} + \epsilon)$ covers the fixed number θ.

2.12.2 Example Calculation: CI

- CI for the mean with normal data and known variance. Let $X \sim N(\mu, \sigma^2)$, then

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

$$Pr(\mu - \epsilon < \bar{X} < \mu + \epsilon) = Pr\left(\frac{\epsilon}{\sigma/\sqrt{n}} < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < \frac{\epsilon}{\sigma/\sqrt{n}}\right)$$

$$= \Phi\left(\frac{\epsilon}{\sigma/\sqrt{n}}\right) - \Phi\left(-\frac{\epsilon}{\sigma/\sqrt{n}}\right)$$

$$= 2\Phi\left(\frac{\epsilon}{\sigma/\sqrt{n}}\right) - 1 = 1 - \alpha$$

- For a given confidence level $1 - \alpha$, we get that

$$\epsilon = z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

where $z_{1-\alpha/2} = \Phi^{-1}(1 - \alpha/2)$.

- So, the following is a $(1 - \alpha)100\%$ CI:

$$\left(\bar{X} - z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}, \bar{X} + z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right)$$

- For a given sample, μ will be either inside or outside this interval. But before drawing the sample, there is a $(1-\alpha)100\%$ chance that an interval constructed this way will cover the true parameter μ.

2.13 Hypothesis Testing

- Main goal is to answer the question:

  > "Given the sample information, is there evidence that the parameter takes a particular value?"

- **Procedure:**
  - State the hypotheses. We will always have two hypotheses that go together:
    - The null hypothesis (denoted H_0) and the alternative hypothesis (denoted H_1).
    - The null hypothesis is the one you want to reject.
  - Gather the evidence:
    - Sample and measure (usually assume random sampling).
  - Compare the evidence to the null hypothesis:
    - Choose and compute the test statistic.
    - Derive the sampling distribution of the statistic under the null.
    - Compute the p-value.
  - Decide whether or not to reject the null hypothesis:
    - Set the level of the test α.
    - Reject the null hypothesis if $p < \alpha$.

- A hypothesis is typically a statement about the population:
  - Null: "The population looks like ..."
  - Alternative: "The population does not look like ..."

- Usually, observed value - expected value = error. We now ask ourselves:

  > "Is this error due to chance? Or something else?"

  - Null: The difference between the sample and the population is due to chance error.
  - Alternative: The difference between the sample and the population is not due to chance error, but to the population being different.

2.13.1 The Errors We Can Make

- We usually reject H_0 if the test statistic is "too large" to be consistent with our null hypothesis:

  > Reject if $|t| > c$ and Do not reject if $|t| \leq c$.

- There are two possible errors we could make:

- Rejecting H_0 when it was really true. This is called a **type I error**.
- Not rejecting H_0 when it was in fact false. This is called a **type II error**.

	H_0 is true	H_0 is false
Not reject H_0	Correct probability $1-\alpha$	Type II error probability β
Reject H_0	Type I error probability α	Correct probability $1-\beta$

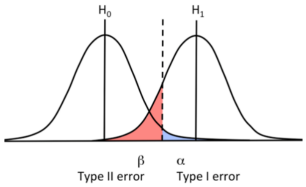


Figure 1: Type 1 and 2 Error

- We want to set c in such a way that it fixes the Type I error rate at an acceptably low level. To compute

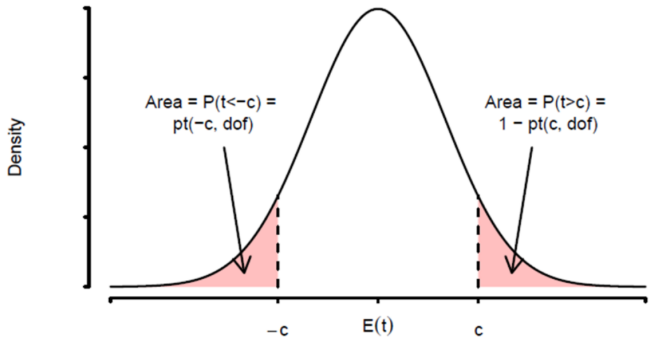
$$Pr(\text{Type I error}) = Pr(|t| > c \mid H_0 \text{ is true}),$$

we need to know the distribution of t under H_0.

  - Then $\alpha \equiv Pr(|t| > c)$, also called the **level of significance**.

We visaulize it like this for a probability distribution:

$$\alpha \equiv \Pr(|t| > c)$$



2.14 CI and Hypothesis Testing

- There is a one-to-one mapping between:

  1. Rejecting H_0 if the statistic exceeds a $\alpha \times 100\%$ critical value.

2. Rejecting H_0 if the hypothesized value of the population parameter lies outside the $(1-\alpha) \times 100\%$ confidence interval (CI).

- Then the point estimate is also "significant at the $\alpha \times 100\%$ level."
- The p-value is the ex-ante probability of observing a t as extreme as in our data.

2.15 Summarized: Estimation, Confidence Interval and Hypothesis Testing

Estimation:

- We start as usual with an assumption about how our data came about (the "data generating process").
- We consider observations in our data $X_1, \ldots, X_n$ as draws from a population or, more generally, an unknown probability distribution $f(X)$.
- The aim is to use our sample data $X_1, \ldots, X_n$ (*remember we assume them to be i.i.d.*) to learn something about the unknown probability distribution $f(X)$ where the data came from.
- We want to estimate a population parameter using the observed data, some measure of variation, an average, min, max, quantile, etc.
- *Point estimation* attempts to obtain a best guess for the value of that parameter.
- An *estimator* is a statistic (function of data) that produces such a guess.

Confidence Intervals:

- We can also quantify the uncertainty (sampling distribution) of our point estimate.
- One way of doing this is by constructing an interval that is likely to contain the population parameter.
- One such interval, which is computed on the basis of the data, is called a *confidence interval*.
- The sampling probability that the confidence interval will indeed contain the parameter value is called the *confidence level*.

Hypothesis Testing:

- Statistical hypothesis testing is a formal method for determining which of the two hypotheses should prevail.
- Each of the two hypotheses predicts a different distribution for the empirical measurements.
- In order to decide which of the distributions is more in tune with the data, a statistic is computed and compared to a threshold for rejection of the hypothesis.
- The threshold is selected in order to assure a small enough probability of erroneously rejecting the currently accepted hypothesis. This is called the *significance level* of the test.

3 Regression Techniques

3.1 Financial Data

- **Cross-Sectional Data:** Data collected at a single point in time, indexed by i. For instance, market cap of various tech companies on a specific date.
- **Time-Series Data:** Data collected over a sequence of time indexed by t. For example, daily prices of Apple stock price over multiple years.
- **Panel Data:** Combination of time-series and cross-sectional data, indexed by it. For example, market cap of various tech companies

For this topic the models are mostly used for cross-sectional data, but is sometimes used for time-series as well.

3.2 Notation:

- Denote the dependent variable by Y and the independent variable(s) by $X_1, ..., X_k$ where there are k independent variables.
- Also, Y_i and $X_{i,1}, ..., X_{i,k}$ are the random values of these variables for the i-the observation.
- Some alternative names for Y and X variables are:
  Y: dependent variable, regressand, effect variable, explained variable.
  X: independent variables, regressors, causal variables, explanatory variables.
- The Y variable is assumed to be stochastic. Hence, there is a probability distribution of Y for each level of X, $f(y|x)$

3.3 Regression in a Nutshell:

- We conduct regression to describe the true relationship between underlying variables.
- Interest is in the probability distribution of Y as a function of X
- **Linear Regression model**: Mean of Y is a straight line function of X.
- **The population regression function (PRF)**:

$$Y = \beta_0 + \beta_1 X + u \quad \text{with} \quad E(u|X = x) = 0,$$

$$E(Y|X = x) = \beta_0 + \beta x$$

- **The sample regression function (SRF)** for a sample $\{(y_i, x_i)\}_{i=1}^n, \quad i = 1, 2, \ldots, n$:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{u}_i$$

- **Goal:** to find good estimates $(\hat{\beta}_0, \hat{\beta}_1)$ for population parameters (β_0, β_1) and to know how well we did.
- Ordinary least squares: we do this by minimizing the squared sum of the error terms in our sample.

3.4 Least-Squares Estimation

- The regression coefficients can be estimated by the method of least squares. The least-squares estimates are the values of $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize:

$$\sum_{i=1}^{n} \left\{ Y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 X_i \right) \right\}^2.$$

- The OLS coefficients of the slope and the intercept are given by:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (Y_i - \bar{Y})(X_i - \bar{X})}{\sum_{i=1}^n (X_i - \bar{X})^2},$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}.$$

- We derive the coefficients by minimizing the function with respect to each of them.

3.5 Simple Regression: An Example

We are provided with the following dataset, which contains:

- **Excess return on the market index** ($R_{M,t} - r_{f,t}$): This represents the market's return over the risk-free rate.
- **Excess return on stock** X ($R_{X,t} - r_{f,t}$): This represents stock X's return over the risk-free rate.

The data for $t = 1, 2, \ldots, 5$ is as follows:

Year (t)	$R_{M,t} - r_{f,t}$	$R_{X,t} - r_{f,t}$
1	13.7	17.8
2	23.2	39.0
3	6.9	12.8
4	16.8	24.2
5	12.3	17.2

The goal is to **fit a linear regression model** that predicts the excess return on stock X (y_t) based on the excess return on the market index (x_t).

The linear regression model can be expressed as:

$$y_t = \beta_0 + \beta_1 x_t + u_t,$$

where:

- β_0: The intercept.
- β_1: The slope (indicates the sensitivity of y_t to changes in x_t).
- u_t: The residual (error term).

The objective is to minimize the **sum of squared residuals**:

$$\min \sum_{t=1}^{T} u_t^2.$$

This gives the best-fit line:

$$\hat{y}_t = \hat{\beta}_0 + \hat{\beta}_1 x_t,$$

where $\hat{\beta}_0$ and $\hat{\beta}_1$ are the estimated coefficients.

Interpretation of the Fitted Line

After performing regression analysis on the given data, we obtain the fitted equation:

$$\hat{y}_t = -1.74 + 1.64 x_t.$$

- **Slope ($\hat{\beta}_1 = 1.64$)**: For every 1-unit increase in the excess return on the market index (x_t), the excess return on stock X (y_t) is expected to increase by 1.64 units. This slope reflects the sensitivity of the stock's return to market movements.
- **Intercept ($\hat{\beta}_0 = -1.74$)**: When the market index's excess return (x_t) is 0, the model predicts that the excess return on stock X (y_t) would be -1.74. This value is often less meaningful unless the data includes values of x_t close to 0.

3.6 Generalizing the Simple Model: Multiple Regression

The dependent variable Y may depend on more than one independent variable X. Hence, we need to include more independent variables to model correctly.

3.6.1 Multiple Regression

- Now we write:

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots + \beta_k X_k + u,$$

  or for $i = 1, 2, \ldots, n$:

$$Y_i = \beta_1 + \beta_2 X_{2,i} + \beta_3 X_{3,i} + \cdots + \beta_k X_{k,i} + u_i.$$

- Now X_1 is the constant term, which is usually represented by a column of ones of length T:

$$X_1 = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}.$$

- **Note**: We have re-labeled the constant term. There are then $(k-1)$ separate independent variables on the RHS of our model, each with its own slope coefficient.
- We could write out a separate equation for every value of t:

$$\begin{aligned} Y_1 &= \beta_1 + \beta_2 X_{2,1} + \beta_3 X_{3,1} + \cdots + \beta_k X_{k,1} + u_1, \\ Y_2 &= \beta_1 + \beta_2 X_{2,2} + \beta_3 X_{3,2} + \cdots + \beta_k X_{k,2} + u_2, \\ &\vdots \\ Y_T &= \beta_1 + \beta_2 X_{2,n} + \beta_3 X_{3,n} + \cdots + \beta_k X_{k,n} + u_n. \end{aligned}$$

- We can write this in matrix form $y = X\beta + u$, where:

$$y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \quad X = \begin{bmatrix} 1 & X_{2,1} & \dots & X_{k,1} \\ 1 & X_{2,2} & \dots & X_{k,2} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{2,n} & \dots & X_{k,n} \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix}, \quad u = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix}.$$

- Dimensions:
  - y is of dimension $n \times 1$,
  - X is of dimension $n \times k$,
  - β is of dimension $k \times 1$,
  - u is of dimension $n \times 1$.

3.6.2 Interpretation of Coefficients

- In the model:

$$Y_i = \beta_1 + \beta_2 X_{2,i} + \beta_3 X_{3,i} + \cdots + \beta_k X_{k,i} + u_i.$$

- Taking the partial derivative of Y_t with respect to any $X_{k,i}$ gives:

$$\frac{\partial E(Y_t \mid X_{2,i} = x_{2,i}, \dots, X_{k,i} = x_{k,i})}{\partial X_{k,i}} = \beta_k.$$

- So we interpret each β_k as the marginal effect of a variation in X_k on Y on average, i.e.,
  - β_k gives us the variation in the expected value of Y given a variation in the value of the k-th regressor *ceteris paribus* (keeping all the other regressors fixed).

- **Polynomial terms**

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \beta_3 X_2 + \cdots + \beta_{k+1} X_k + u$$

  - The effect of a change in X_1 is $\beta_1 + 2\beta_2 X_1$ *ceteris paribus*.
  - Example:

$$\text{StockPrice}_i = 5 + 1.5\,\text{YearsSinceIPO}_i - 0.5\,\text{YearsSinceIPO}_i^2 + u_i.$$

- **Log-Transformed Variables**

$$Y = \beta_0 + \beta_1 \ln(X_1) + \beta_2 X_2 + \cdots + \beta_k X_k + u$$

  - A percentage change in X_1 leads to a $\frac{\beta_1}{100}$ unit change in Y *ceteris paribus*.

- $$\ln(Y) = \beta_0 + \beta_1 \ln(X_1) + \beta_2 X_2 + \cdots + \beta_k X_k + u$$
  - A percentage change in X_1 leads to a β_1 percent change in Y *ceteris paribus*.
- $$\ln(Y) = \beta_0 + \beta_1 \ln(X_1) + \beta_2 X_2 + \cdots + \beta_k X_k + u$$
  - A unit change in X_2 leads to a β_2 percent change in Y *ceteris paribus*.
- **Note:** When using logs we interpret things in percentage change.

3.6.3 Dummy Variables

- Dummy variables take only value 1 or 0, where zero always refers to the reference group for each dummy-coded predictor.
- They help us to identify special effects.
  - For example, for the sample of students in our class, define a variable that takes value 1 if the i-th student has taken previously an econometrics course, zero if not.

$$Y = \beta_0 + \beta_1 D_1 + \beta_2 X_1 + \cdots + \beta_{k+1} X_k + u$$

- In this model, the intercept β_0 represents the mean of Y for the reference group ($D_1 = 0$) when all continuous predictors $X_1, \ldots, X_k$ are zero, *ceteris paribus*.
  - So this is an example where the intercept becomes meaningful and useful for answering hypotheses.
- The coefficient β_1 represents the difference in the mean of Y between the comparison group ($D_1 = 1$) and the reference group ($D_1 = 0$), *ceteris paribus*.
  - Example:

    $$\text{StockReturn}_t = 0.03 + 0.5\,\text{GoodNewsDayDummy}_t + 1.2\,\text{MarketReturn}_t + u_t$$

- Dummy variables can capture effects on the slope coefficients too.
  - Example:

    $$\text{StockReturn}_t = 0.03 + 1.2\,\text{MarketReturn}_t + 0.2\,\text{GoodNewsDayDummy}_t \times \text{MarketReturn}_t + u_t$$

3.6.4 Dummy Variable Trap

- A possible issue while using dummy variables is the **dummy variable trap**.
- The dummy variable trap arises when the incorrect use of dummies generates perfect multicollinearity in a regression model:
  - i.e., it makes one of your explanatory variables a perfect linear combination of some others such that rank(X) is no longer full.
  - **Note:** You can introduce as many dummies as you need as long as you are not using the intercept and **all** the dummies.
- Suppose that we are assessing if there is a day-of-the-week effect in a financial market.
- There are only 5 trading days in a week, so you can create 5 dummies:

$$r_t = \beta_0 + \beta_1 r_t^M + \beta_2 D_{1t} + \cdots + \beta_6 D_{5t} + u_t$$

$$D_{5t} = 1 - (D_{1t} + D_{2t} + D_{3t} + D_{4t}) \quad \Rightarrow \quad X'X \text{ cannot be inverted!}$$

- **Note:** In this course X' refers to the transposed of X! Usually we use X^T
- **Solution:**
  - Drop the intercept:

$$r_t = \beta_1 r_t^M + \beta_2 D_{1t} + \cdots + \beta_6 D_{5t} + u_t$$

    and interpret the dummies' parameters as the intercept for Monday, Tuesday, and so on.
  - **OR** drop any of the dummies:

$$r_t = \beta_0 + \beta_1 r_t^M + \beta_2 D_{1t} + \cdots + \beta_5 D_{4t} + u_t$$

    where β_0 is the intercept on Friday, and the other coefficients represent the difference between Fridays and other days.

3.7 Estimation: Multiple Regression Model

3.7.1 The OLS Estimator: The Estimator of the Parameter Vector

- Derivations of the β's follow exactly the same steps as before.
- We minimize the sum of squared residuals through the choice of parameter vector β.
- The fitted model is:

$$\hat{y} = X\hat{\beta}.$$

- The $n \times 1$ vector of residuals:

$$\hat{u} = y - \hat{y} = y - X\hat{\beta}.$$

- The sum of squared residuals:

$$\hat{u}'\hat{u} = \sum_i \hat{u}_i^2.$$

- Residuals in matrix form:

$$\hat{u}'\hat{u} = \begin{bmatrix} \hat{u}_1 \\ \hat{u}_2 \\ \vdots \\ \hat{u}_n \end{bmatrix}^T \begin{bmatrix} \hat{u}_1 \\ \hat{u}_2 \\ \vdots \\ \hat{u}_n \end{bmatrix} = [\hat{u}_1\, \hat{u}_2\, \ldots\, \hat{u}_n] \begin{bmatrix} \hat{u}_1 \\ \hat{u}_2 \\ \vdots \\ \hat{u}_n \end{bmatrix}.$$

  This results in:

$$\hat{u}_1^2 + \hat{u}_2^2 + \cdots + \hat{u}_n^2 = \sum_i \hat{u}_i^2.$$

- In order to obtain the parameter estimates, $\beta_1, \beta_2, \ldots, \beta_k$, we would minimize the RSS (Residual Sum of Squares) with respect to all the β's.

- It can be shown that:

$$\hat{\beta} = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_k \end{bmatrix} = (X'X)^{-1}X'y.$$

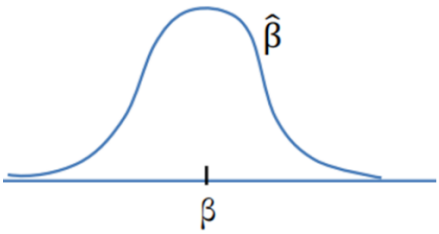
- Check the dimensions: $k \times 1$.

3.7.2 Distribution of OLS estimates

- If we believe in a structural form and run a regression (OLS), the output from the model is:
  - Estimates for $\beta_1, \beta_2, \ldots, \beta_k$: $\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_k$,
  - Fitted values for $y, \hat{y}$, and residuals, $\hat{u}$.

- **We need to know how good our estimates are**, that is:
  - We want to say something about β based on $\hat{\beta}_2$.
  - $\hat{\beta}_2$ will take different values when we take a new sample.
  - Hence, $\hat{\beta}_2$ is a random variable and has a probability distribution.
  - Based on our value for $\hat{\beta}_2$, we can calculate probabilities for different values of β_2.

– **For this, we need to make assumptions!**



3.7.3 The Assumptions of Classical Linear Regression Model (CLRM)

We have the following assumptions:

1. $E(u_i) = 0$
2. $Var(u_i) = \sigma^2 < \infty$
3. $Cov(u_i, u_j) = 0$
4. $Cov(u_i, X_i) = 0$
5. $u \sim \mathcal{N}(0, \sigma^2)$
6. Assumption of linearity

Note that X is assumed to be a random variable.

3.7.4 Properties of OLS Estimator

If assumptions 1 to 4 hold, then the estimators $\hat{\beta}$ determined by OLS are known as Best Linear Unbiased Estimators (BLUE).

- **Estimator** - $\hat{\beta}$ is an estimator of the true value of β
- **Linear** - $\hat{\beta}$ is a linear estimator
- **Unbiased** - on average, the actual value of the $\hat{\beta}$'s will be equal to the true values
- **Best** - means that the OLS estimator has the minimum variance among the class of linear unbiased estimators

3.7.5 Precision and Standard Errors

- What we need is some measure of the reliability or precision of the estimators.
- To be able to say something (meaningful) about β, we need to make assumptions about our estimate's probability distribution:
  - We want unbiasedness: $E[\hat{\beta}] = \beta$, and we need the correct value for $\text{Var}[\hat{\beta}]$, we usually use:

$$\sqrt{\text{Var}[\hat{\beta}]} = SE(\hat{\beta}).$$

- Given assumptions 1–4 above, the parameter variance-covariance matrix is given by:

$$\text{Var}(\hat{\beta}) = \sigma^2(X'X)^{-1}.$$

- The problem that we now encounter is that the (conditional) distributions of the estimates depend on the standard deviation of the error term σ.
- The variance of the random variable u_t is given by:

$$\text{Var}(u_i) = E[(u_i - E(u_i))^2] = E(u_i^2) = \sigma^2.$$

- This parameter is, typically, not known, moreover u_t cannot be observed. Therefore, we have to replace σ by an unbiased estimator, s:

$$s^2 = \frac{\hat{u}'\hat{u}}{T-k},$$

  where k = number of regressors.
- The parameter variance-covariance matrix is given by:

$$\text{Var}(\hat{\beta}) = s^2(X'X)^{-1}.$$

3.8 Inference in Regression: Hypothesis Testing

We want to make inferences about the likely population values from the regression parameters. If we know the distribution of of our estimate we can say something about the population value (loosely speaking). For this we use tools from hypothesis testing.

3.8.1 Hypothesis Testing

- We assume that:

$$u_t \sim \mathcal{N}(0, \sigma^2).$$

- Since the least squares estimators are linear combinations of the random variables, the weighted sum of normal random variables is also normally distributed:
$$\hat{\beta}_k \sim \mathcal{N}(\beta_k, \text{Var}(\hat{\beta}_k)).$$

- Constructing standard normal variables and replacing standard deviations with their sample estimates:
$$\frac{\hat{\beta}_k - \beta_k}{\sqrt{\text{Var}(\hat{\beta}_k)}} \sim \mathcal{N}(0,1) \quad \Rightarrow \quad \frac{\hat{\beta}_k - \beta_k}{\sqrt{\widehat{\text{Var}}(\hat{\beta}_k)}} = \frac{\hat{\beta}_k - \beta_k}{SE(\hat{\beta}_k)} \sim t_{n-k}.$$

3.8.2 The Test of Significance Approach: T-test

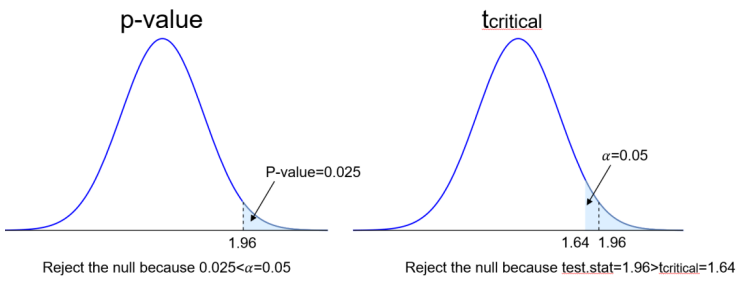
The steps involved in doing a test of significance are:

1. Estimate $\hat{\beta}_k$ and $SE(\hat{\beta}_k)$ in the usual way
2. Calculate the test statistic, where β_k^* is the value of (for example) β_k
$$\text{test stat.} = \frac{\hat{\beta}_k - \beta_k^*}{SE(\hat{\beta}_k)}$$
3. Determine the proper distribution. In this case t-distribution with $n-k$ degrees of freedom.
4. Choose a significance level, often denoted α
5. Determine the rejection region
6. Use the t-tables to obtain a critical value or values with which to compare the test statistic
7. Finally perform the test. If the test statistic lies in the rejection region then reject the null hypothesis (H_0), else do not reject H_0.

3.8.3 P-value and T-critical

Remember that we can also perform hypothesis tests based on the p-value. We compare the two methods visually, to understand the difference:

- Suppose test stat $= 1.96$ and $\alpha = 0.05$. Then we reject the null hypothesis for these values:



3.8.4 CI Approach to Hypothesis Testing

- An example of its usage: We estimate a parameter, say β_k, to be 0.93, and a "95% confidence interval" to be (0.77, 1.09). This means that we are 95% confident that the interval contains the true (but unknown) value of β.
- Confidence intervals are almost invariably two-sided, although in theory, a one-sided interval can be constructed.
- CI approach and test of significance approach will always give the same result.
- **How to Carry out a Hypothesis Test**
  1. Calculate $\hat{\beta}_k$ and $SE(\hat{\beta}_k)$, as before.
  2. Choose a significance level, α (again the convention is 5%). This is equivalent to choosing a $(1-\alpha)\times 100\%$ confidence interval, i.e., 5% significance level = 95% confidence interval.
  3. Use the t-tables to find the appropriate critical value, which will again have $n-k$ degrees of freedom.
  4. The confidence interval is given by:

$$\left(\hat{\beta}_k - t_{\text{crit}} \times SE(\hat{\beta}_k), \hat{\beta}_k + t_{\text{crit}} \times SE(\hat{\beta}_k)\right).$$

  5. Perform the test: If the hypothesized value of β_k (β_k^*) lies outside the confidence interval, then reject the null; otherwise, do not.

3.8.5 A Special Type of Hypothesis Test: The t-ratio

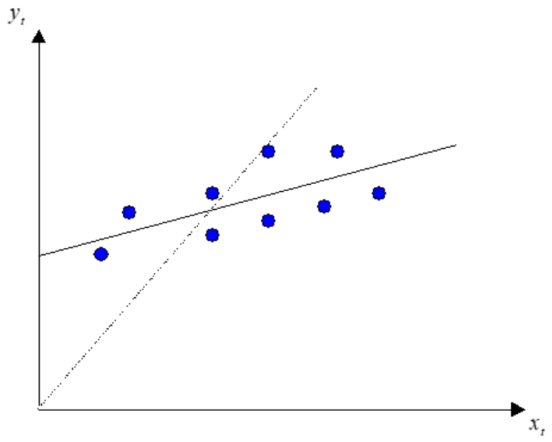
- Recall that the formula for a test of significance approach to hypothesis testing using a t-test was:

$$\text{test statistic} = \frac{\hat{\beta}_k - \beta_k^*}{SE(\hat{\beta}_k)}.$$

- If the test is:
$$H_0 : \beta_k = 0, \quad H_1 : \beta_k \neq 0.$$
- Since $\beta_k^* = 0$,
$$\text{test stat} = \frac{\hat{\beta}_k}{SE(\hat{\beta}_k)}.$$
- The ratio of the coefficient to its SE is known as the *t-ratio* or *t-statistic*.

3.8.6 What the t-ratio Tells Us

If we reject H_0, we say that the result is significant. If the coefficient is not "significant" (e.g. the intercept coefficient in the regression), then it means that the variable is not helping to explain variations in y. In practice, however, there are good statistical reasons for always having a constant even if it is not significant. We illustrate this by showing what happens to a linear regression if the intercept term is not included:



3.8.7 P-value Quick Recap

We remember that, our sample statistic t is a random variable and that the p-value is the ex-ante probability of observing a t as extreme as in our data.

3.9 Inference in Regression: Testing Multiple Hypotheses

3.9.1 The F-test

We test multiple hypotheses by using something called the F-test.

- The F-test involves estimating 2 regressions:
  - The **unrestricted regression** is the one in which the coefficients are freely determined by the data, as we have done before.

- The **restricted regression** is the one in which the coefficients are restricted, i.e., the restrictions are imposed on some β's.

- **Example**
  - The general regression is:

$$Y_i = \beta_1 + \beta_2 X_{2,i} + \beta_3 X_{3,i} + \beta_4 X_{4,i} + u_i$$

  - We want to test the restriction that $\beta_3 + \beta_4 = 1$. The unrestricted regression is the equation above, but what is the restricted regression?
  - We substitute the restriction $\beta_3 + \beta_4 = 1$ into the regression so that it is automatically imposed on the data:

$$\beta_3 + \beta_4 = 1 \implies \beta_4 = 1 - \beta_3$$

$$Y_i = \beta_1 + \beta_2 X_{2,i} + \beta_3 X_{3,i} + (1 - \beta_3) X_{4,i} + u_i$$

  - Gather terms in β's together to get the restricted regression:

$$(Y_i - X_{4,i}) = \beta_1 + \beta_2 X_{2,i} + \beta_3 (X_{3,i} - X_{4,i}) + u_i$$

  - We create two new variables, $P_t = Y_i - X_{4,i}$ and $Q_t = X_{3,i} - X_{4,i}$.
  - The restricted regression we actually estimate is:

$$P_i = \beta_1 + \beta_2 X_{2,i} + \beta_3 Q_i + u_i.$$

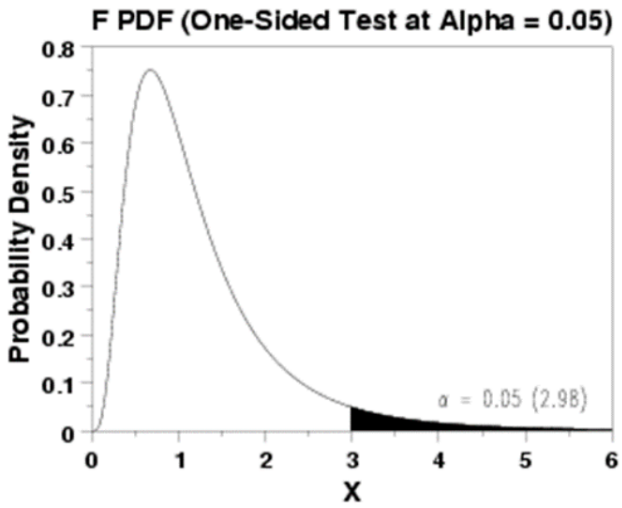
- The test statistic is given by:

$$\text{test statistic} = \frac{\text{RRSS} - \text{URSS}}{\text{URSS}} \times \frac{n-k}{m},$$

  where:

  - URSS = RSS from unrestricted regression,
  - RRSS = RSS from restricted regression,
  - m = number of restrictions,
  - n = number of observations,
  - k = number of regressors in unrestricted regression including a constant in the unrestricted regression (or the total number of parameters to be estimated).

3.9.2 The F-distribution



- The test statistic follows the F-distribution, which has 2 degrees of freedom (d.f.) parameters:
  - The value of the degrees of freedom parameters are m and $(n-k)$, respectively (the order of the d.f. parameters is important).
  - The appropriate critical value will be in column m, row $(n-k)$.
  - The F-distribution has only positive values and is not symmetrical.
- **Rejection region**: We reject the restrictions if the value of the test statistic is larger than the critical value from the F-distribution, based on the chosen significance level, i.e.,

$$\text{the test statistic} > \text{critical F-value}.$$

  - Note that this is by definition a two-sided test always, but we reject only for very large values of the test statistic, so we reject on the right tail.
  - The test statistic is positive by definition since RRSS is always bigger than the URSS from constrained optimization using the OLS.
  - So we reject the two-sided alternative for large values of the test statistic. The F-test is not flexible enough to allow for a one-sided alternative hypothesis.

3.9.3 F-test Intuition

- The unrestricted regression will always fit the data at least as well as the restricted regression.

- Only when the restriction is perfectly valid, though, will the restricted regression fit just as well as the unrestricted.
  - In this case, RRSS is approximately equal to URSS, so the F-statistic will be approximately zero.
- If the restriction is completely false in the data, the restricted regression will fit much worse than the unrestricted.
  - In this case, RRSS > URSS, so the F-statistic will be big and positive.
  - So, when the F-statistic is big, that indicates that the null hypothesis should be rejected, and this is exactly what happens.

3.9.4 Example: No. of restrictions, m

H_0 : hypothesis	
$\beta_1 + \beta_2 = 2$	1
$\beta_2 = 1$ and $\beta_3 = -1$	2
$\beta_2 = 0, \beta_3 = 0$ and $\beta_4 = 0$	3

- If the model is $Y_i = \beta_1 + \beta_2 X_{2,i} + \beta_3 X_{3,i} + \beta_4 X_{4,i} + u_i$, then the null hypothesis $H_0 : \beta_2 = 0, \beta_3 = 0$ and $\beta_4 = 0$ is tested by the regression F-statistic.
- Note the form of the alternative hypothesis for all tests when more than one restriction is involved: $H_1 : \beta_2 \neq 0$, or $\beta_3 \neq 0$, or $\beta_4 \neq 0$.

3.10 Goodness of Fit Statistics

- We would like some measure of how well our regression model actually fits the data.
- We have goodness of fit statistics to test this.
- The most common goodness of fit statistic is known as R^2.
  - In simple terms, it is the square of the correlation coefficient between y and $\hat{y}$.
- This measure tells us how much of the data variability is explained by the model.
- Formally: we are interested in explaining the variability of y about its mean value, i.e., the total sum of squares (TSS):

$$TSS = \sum_t (y_i - \bar{y})^2$$

- We can split the TSS into two parts:

  - The part which we have explained (known as the explained sum of squares, ESS)
    * The Explained Sum of Squares (ESS), equal to $\sum_i (\hat{y}_i - \bar{y})^2$, measures the variation in the predicted values of the regression around the mean of y.
  - And the part which we did not explain using the model (the RSS).
    * The Residual Sum of Squares (RSS), equal to $\sum_i \hat{u}_i = \sum_t (y_i - \hat{y}_i)^2$, is the part of the variation in y that the regression can't explain.

- So we have:

$$TSS = ESS + RSS$$

3.10.1 Defining R^2

We define $R^2 = \frac{ESS}{TSS}$

3.10.2 Problems with R^2

- R^2 is defined in terms of variation about the mean of y so that if a model is reparametrized (rearranged) and the dependent variable changes, R^2 will change.
- R^2 never falls if more regressors are added to the regression, e.g., consider:
  - Regression 1: $Y_i = \beta_1 + \beta_2 X_{2,i} + \beta_3 X_{3,i} + u_i$
  - Regression 2: $Y_i = \beta_1 + \beta_2 X_{2,i} + \beta_3 X_{3,i} + \beta_4 X_{4,i} + u_i$
- R^2 will always be at least as high for Regression 2 relative to Regression 1.
- High R^2 might be due to overfitting.

3.10.3 R^2 and its Goal

- If your main goal is to determine which variables are statistically significant and how changes in the independent variables relate to changes in the dependent variable, R^2 is almost totally irrelevant.
  - If you correctly specify a regression model, the R^2 value doesn't affect how you interpret this relationship.
- If your main goal is to produce precise predictions, R^2 becomes a concern.
  - A low R^2 indicates that your model doesn't explain much of the variance in the data. However, it doesn't necessarily mean that your predictions are imprecise.
  - A high R^2 might occur because the model fits noise in the data, leading to poor performance on unseen data.

3.10.4 Alternatives to R^2

- Adjusted R^2: a modification which takes into account the loss of degrees of freedom associated with adding extra variables.

$$\bar{R}^2 = 1 - \left[\frac{n-1}{n-k}(1-R^2)\right]$$

- So if we add an extra regressor, k increases, and unless R^2 increases by a more than offsetting amount, it will actually fall.
- Predicted R^2 can be used to determine how well a regression model makes predictions and tackles the overfitting problem:
  - Software calculates it as follows:
    1. Removes a data point from the dataset.
    2. Calculates the regression equation.
    3. Evaluates how well the model predicts the missing observation.
    4. Repeats this for all data points.
- There are still problems with the criterion:
  - A "soft" rule (leads to a very large model).
  - No distribution for R^2 or $\bar{R}^2$.

3.11 Hypothesis Testing Under Maximum Likelihood

Hypothesis testing based on maximum likelihood estimation (MLE) provides a more general framework compared to traditional tests such as F-tests and t-tests, especially for non-linear models. Three primary hypothesis testing procedures based on maximum likelihood principles are:

- Wald test,
- Likelihood Ratio (LR) test,
- Lagrange Multiplier (LM) test.

These tests evaluate the degree to which the maximal value of the likelihood function (LLF) falls upon imposing a restriction. If the LLF falls significantly, the restriction is deemed unsupported by the data, leading to the rejection of the hypothesis.

3.11.1 Graphical Interpretation of the Tests

Consider a single parameter θ to be estimated, with $L(\theta)$ as the log-likelihood function. Let $\hat{\theta}$ denote the unrestricted maximum likelihood estimate (MLE) and $\tilde{\theta}$ the restricted estimate. The three tests differ in how they assess the effect of the restriction:

- The **Likelihood Ratio (LR) test** is based on the **vertical distance** between the likelihood values $L(\hat{\theta})$ and $L(\tilde{\theta})$. This measures the drop in the log-likelihood due to the restriction.
- The **Wald test** is based on the **horizontal distance** between $\hat{\theta}$ and $\tilde{\theta}$. This measures how far the unrestricted estimate deviates from the restricted one.
- The **Lagrange Multiplier (LM) test** compares the **slopes** of the likelihood function at $\tilde{\theta}$ and $\hat{\theta}$. The unrestricted MLE $L(\hat{\theta})$ has a zero slope, and the LM test checks if the slope at $L(\tilde{\theta})$ is significantly steep.

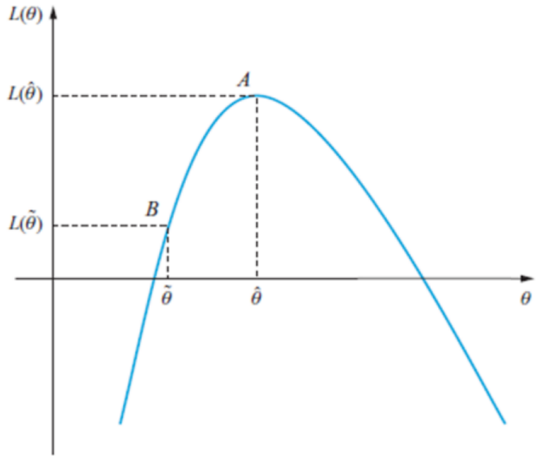


Figure 2: Illustration of the Wald, LR, and LM tests in the context of MLE.

Each of these tests provides a different perspective on whether the imposed restriction significantly affects the likelihood, making them useful for different scenarios in hypothesis testing under MLE.

3.11.2 Distributions for Diagnostic Tests

Regression diagnostic tests rely on the calculation of test statistics, which follow specific distributions. The choice of the test statistic and its distribution depend on how the hypothesis test is constructed.

Distribution of Test Statistics:

- The **LM test** statistic follows a chi-squared (χ^2_m) distribution, where m is the number of restrictions imposed on the model.
- The **Wald test** statistic follows an F-distribution with $(m, T-k)$ degrees of freedom, where T is the sample size and k represents the number of estimated parameters.
- Asymptotically, the Wald test and the LM test become equivalent because the χ^2 distribution is a special case of the F-distribution.
- For small samples, the F-distribution version of the Wald test is preferable due to its better finite-sample properties.

Relationship Between Distributions:

Understanding the connections between different distributions used in hypothesis testing is crucial for selecting the appropriate test. The relationships between standard normal, chi-squared, and t-distributions lead to the F-distribution:

- Let $Z \sim N(0,1)$, then:
  - $Z^2 \sim \chi^2_1$ (a standard normal squared follows a chi-squared distribution with 1 degree of freedom).
  - The sum of k independent squared standard normal variables follows a χ^2_k distribution: $Z_1^2 + Z_2^2 + \cdots + Z_k^2 \sim \chi^2_k$.
  - The ratio of a standard normal variable to the square root of a chi-squared variable divided by its degrees of freedom follows a t-distribution:
$$\frac{Z}{\sqrt{\chi^2_n/n}} \sim t_n.$$
- The F-distribution emerges from the ratio of two independent chi-squared variables normalized by their degrees of freedom:
$$F(m, T-k) \sim \frac{\chi^2_m/m}{\chi^2_{T-k}/(T-k)}.$$
- As the sample size n increases, the F-distribution and chi-squared distribution converge:
  - $F(m,n) \to \frac{\chi^2_m}{m}$ as $n \to \infty$.
  - The squared t-distribution follows an F-distribution with 1 and n degrees of freedom: $t_n^2 \sim F(1,n)$.

This framework helps in understanding the underlying assumptions and properties of hypothesis testing procedures, ensuring the correct application of statistical tests based on the distribution of the test statistic.

3.12 The Gauss-Markov Assumptions

3.12.1 Assumptions of the Classical Linear Regression Model (CLRM)

The Classical Linear Regression Model (CLRM) relies on a set of key assumptions regarding the disturbance terms u_t. These assumptions are essential for the ordinary least squares (OLS) estimator to be unbiased and efficient. The standard assumptions are as follows:

1. **Zero Mean of Errors**: $E(u_t) = 0$.
2. **Homoscedasticity**: $\text{Var}(u_t) = \sigma^2 < \infty$.
3. **No Autocorrelation**: $\text{Cov}(u_i, u_j) = 0$ for all $i \neq j$.
4. **Exogeneity of Regressors**: The regressor matrix X is non-stochastic or fixed in repeated samples, or equivalently, $\text{Cov}(u_t, x_t) = 0$.
5. **Normality of Errors** (optional for small sample inference): $u_t \sim N(0, \sigma^2)$.
6. **No Perfect Multicollinearity**: None of the regressors is a perfect linear combination of the others.

Instead of assuming both Assumption 1 and Assumption 4, we impose the stronger condition:

$$E(u_t|x_t) = 0.$$

This assumption ensures that the disturbance term u_t has an expected value of zero for every value of x_t, rather than just on average. The key reasons for making this substitution are:

- The condition $E(u_t) = 0$ alone does not prevent the errors from being correlated with the regressors.
- The assumption $\text{Cov}(u_t, x_t) = 0$ only ensures that errors and regressors are uncorrelated *linearly*, but it does not rule out nonlinear dependencies.
- The stronger assumption $E(u_t|x_t) = 0$ guarantees that u_t is mean independent of x_t, which is sufficient to ensure that the OLS estimator remains unbiased and consistent.

Thus, replacing Assumptions 1 and 4 with the single assumption $E(u_t|x_t) = 0$ simplifies the exogeneity requirement while strengthening its implications for unbiased estimation in the CLRM framework.

We will now study each of these assumptions further.

3.12.2 Bias: Assumption 1

This assumptions states that the mean of disturbances is zero. For all diagnostic tests, we cannot observe the disturbances and so perform the test of residuals. We know that the mean of residuals will always be zero provided that there is a constant term in the regression.

3.12.3 Heteroscedasticity: Assumptions 2

Assumption 2 states that the variance is constant for the errors, which is known as homoscedasticity. If the errors do not have constant variance, they are known as heteroscedastic.

Detection of Heteroscedasticity There are two main approaches to detect heteroscedasticity:

- **Graphical Methods**: Examining scatter plots of residuals against fitted values or independent variables to visually inspect patterns of changing variance.
- **Formal Statistical Tests**: These include White's test, and the ARCH test.

The Goldfeld-Quandt (GQ) Test The Goldfeld-Quandt test is one of the classical methods to detect heteroscedasticity. It involves splitting the dataset and comparing the variances of residuals across the two subsets.

Steps for the Goldfeld-Quandt test:

1. Split the total sample (of size T) into two sub-samples of sizes T_1 and T_2, ensuring that they are ordered according to an explanatory variable.
2. Estimate the regression separately on each sub-sample.
3. Compute the residual variances S_1^2 and S_2^2 for both groups.
4. Compute the test statistic:
$$GQ = \frac{S_1^2}{S_2^2}.$$
5. Under the null hypothesis of homoscedasticity, this statistic follows an F-distribution:
$$F(T_1 - k, T_2 - k).$$
6.

Limitations:

- The choice of where to split the sample is arbitrary and may significantly affect test results.

White's Test for Heteroscedasticity White's test is a more general approach to detecting heteroscedasticity without assuming a specific pattern. Instead of splitting the sample, it examines whether the variance of residuals depends on the independent variables.

Steps for White's Test:

1. Estimate the original regression model:

$$y_t = \beta_1 + \beta_2 x_{2t} + \beta_3 x_{3t} + u_t.$$

2. Obtain the squared residuals $\hat{u}_t^2$.

3. Run the auxiliary regression:

$$\hat{u}_t^2 = \alpha_1 + \alpha_2 x_{2t} + \alpha_3 x_{3t} + \alpha_4 x_{2t}^2 + \alpha_5 x_{3t}^2 + \alpha_6 x_{2t} x_{3t} + v_t.$$

4. Perform one of the following significance tests:

   - **F-test** for overall significance of the auxiliary regression.
   - **LM test**: The test statistic is computed as:

$$TR^2 \sim \chi^2(m),$$

     where m is the number of regressors in the auxiliary regression excluding the constant.

5. If the χ^2 test statistic exceeds the critical value, we reject the null hypothesis of homoscedasticity.

Limitations:

- White's test does not specify whether heteroscedasticity depends on time or the order of observations.
- It does not account for time series features like autocorrelation or volatility clustering.

ARCH-test:
This is used for time-dependent heteroscedasticity, which will be covered in the time-series part of this course.

Consequences of Heteroscedasticity:

- OLS estimation still gives unbiased coefficient estimates, but they are no longer BLUE
- This implies that if we still use OLS in the presence of heteroscedasticity, our standard errors could be inappropriate and hence any inferences we make could be misleading.

- Whether the standard errors calculated using the usual formulae are too big or too small will depend upon the form of the heteroscedasticity.

How to Combat Heteroscedaticity:
Generalized Least Squares (GLS):
If the form of heteroscedasticity is known, we can apply the *Generalized Least Squares (GLS)* method to transform the regression equation so that the new error term has constant variance.

Assume the error variance is related to another variable z_t, such that:

$$\text{Var}(u_t) = \sigma^2 z_t^2.$$

To correct for heteroscedasticity, we divide the regression equation by z_t:

$$\frac{y_t}{z_t} = \beta_1 \frac{1}{z_t} + \beta_2 \frac{x_{2t}}{z_t} + \beta_3 \frac{x_{3t}}{z_t} + v_t,$$

where the new error term is:

$$v_t = \frac{u_t}{z_t}.$$

Since the variance of u_t was $\sigma^2 z_t^2$, the transformed variance of v_t becomes:

$$\text{Var}(v_t) = \text{Var}\left(\frac{u_t}{z_t}\right) = \frac{\sigma^2 z_t^2}{z_t^2} = \sigma^2.$$

This transformation eliminates heteroscedasticity, making OLS valid again.

Other Approaches:
When the exact form of heteroscedasticity is unknown, alternative solutions include:

- **Variable Transformations:** Converting variables into logarithmic form or scaling them using a measure of "size" (e.g., dividing income by population for per capita income).
- **White's Heteroscedasticity-Consistent Standard Errors:** A widely used solution that adjusts standard errors without transforming the model.

White's Heteroscedasticity-Consistent Standard Errors
White's method corrects the standard errors of OLS coefficients so that they remain valid even in the presence of heteroscedasticity.

- The corrected standard errors are typically larger than the uncorrected OLS errors.
- This adjustment makes hypothesis tests more **conservative**, reducing the likelihood of falsely rejecting the null hypothesis.

3.12.4 Autocorrelation: Assumption 3

This is essentially the same as saying there is no pattern in the errors. Obviously, we never have the actual u's, so we use their sample counterpart, the residuals. If there are patterns in the residuals from a model, we say that they are autocorrelated.

Detecting Autocorrelation: Durbin-Watson Test:
The Durbin-Watson (DW) test is a statistical test for detecting first-order autocorrelation in the residuals of a regression model. The test examines whether adjacent residuals are correlated and is based on the model:

$$u_t = \rho u_{t-1} + v_t, \quad v_t \sim N(0, \sigma_v^2)$$

where ρ is the autocorrelation coefficient. The null hypothesis (H_0) and alternative hypothesis (H_1) for the test are:

$$\begin{aligned} H_0 &: \rho = 0 \quad \text{(No autocorrelation)} \\ H_1 &: \rho \neq 0 \quad \text{(Autocorrelation present)} \end{aligned}$$

The test statistic is computed as:

$$DW = \frac{\sum_{t=2}^{T} (\hat{u}_t - \hat{u}_{t-1})^2}{\sum_{t=2}^{T} \hat{u}_t^2} \approx 2(1 - \hat{\rho})$$

where $\hat{\rho}$ is the estimated sample correlation between $\hat{u}_t$ and $\hat{u}_{t-1}$.

Since $\hat{\rho}$ is constrained between -1 and 1, the Durbin-Watson statistic satisfies:

$$0 \leq DW \leq 4.$$

The interpretation of DW is as follows:

- If $\hat{\rho} = 0$, then $DW \approx 2$, indicating little to no autocorrelation.
- If $DW < 2$, there is evidence of positive autocorrelation.
- If $DW > 2$, there is evidence of negative autocorrelation.

The test utilizes two critical values, a lower bound (d_L) and an upper bound (d_U), which depend on the sample size and number of regressors. The decision rule is:

- If $DW < d_L$, reject H_0 in favor of positive autocorrelation.
- If $DW > 4 - d_L$, reject H_0 in favor of negative autocorrelation.
- If $d_L \leq DW \leq d_U$, the test is inconclusive.

- If $d_U < DW < 4 - d_U$, do not reject H_0, indicating no significant autocorrelation.

Other Tests for Autocorrelation:
Several other tests for autocorrelation exist, including:

- **Breusch-Godfrey test**: A general test for higher-order autocorrelation.
- **Box-Pierce test**: A test based on the sum of squared autocorrelations for simpler model structures.
- **Box-Ljung test**: Similar to the Box-Pierce test but includes a correction factor for small sample sizes.

Breusch-Godfrey Test The Breusch-Godfrey test is a more general test for detecting autocorrelation of order r. The test involves estimating the following auxiliary regression:

$$u_t = \rho_1 u_{t-1} + \rho_2 u_{t-2} + \cdots + \rho_r u_{t-r} + v_t.$$

The null and alternative hypotheses for the test are:

$$H_0 : \rho_1 = 0, \rho_2 = 0, \ldots, \rho_r = 0 \quad \text{(No autocorrelation)}$$
$$H_1 : \rho_1 \neq 0 \text{ or } \rho_2 \neq 0 \text{ or } \ldots \text{ or } \rho_r \neq 0 \quad \text{(Autocorrelation present)}$$

The test procedure consists of the following steps:

1. Estimate the original regression model using ordinary least squares (OLS) and obtain the residuals $\hat{u}_t$.
2. Regress $\hat{u}_t$ on all the regressors from the original model along with lagged values $\hat{u}_{t-1}, \hat{u}_{t-2}, \ldots, \hat{u}_{t-r}$.
3. Obtain the coefficient of determination R^2 from this regression.
4. Compute the test statistic:

$$(T - r)R^2 \sim \chi^2(r),$$

where T is the sample size and r is the order of autocorrelation tested.

If the computed test statistic exceeds the critical value from the chi-squared distribution, reject H_0, indicating the presence of autocorrelation.

Consequences of Ignoring Autocorrelation:
The coefficient estimates derived using OLS are still unbiased, but they are inefficient, i.e. they are not BLUE, even in large sample sizes. Thus, if the standard error estimates are inappropriate, there exists the possibility that we could make the wrong inferences. R^2 is likely to be inflated relative to its "correct" value

for positively correlated residuals.

Addressing Autocorrelation:
If the form of autocorrelation is known, one possible approach is to use a Generalized Least Squares (GLS) procedure, such as the Cochrane-Orcutt method, which explicitly accounts for autocorrelated residuals. However, these methods require assumptions about the nature of the autocorrelation. If these assumptions are incorrect, they can lead to model misspecification.
Since the precise form of autocorrelation is often unknown, an alternative approach is to modify the regression model by incorporating dynamic terms:

$$y_t = \beta_1 + \beta_2 x_{2t} + \beta_3 x_{3t} + u_t.$$

One can introduce lagged values of both the dependent and independent variables:

$$y_t = \beta_1 + \beta_2 x_{2t} + \beta_3 x_{3t} + \gamma y_{t-1} + \beta_2 x_{2,t-1} + \beta_3 x_{3,t-1} + u_t.$$

Including additional lags can help address autocorrelation, especially in cases where:

- The dependent variable exhibits natural correlation over time.
- An omitted variable, which is itself autocorrelated, influences the regression.
- The dependent variable follows a seasonal pattern that has not been explicitly modeled.

Nevertheless, any modifications must be theoretically justified, as adding lagged terms can introduce biases and overfitting issues.

3.12.5 Endogenity: Assumption 4

In the context of Ordinary Least Squares (OLS) regression, an essential assumption is that the regressors are exogenous. This is formally expressed as:

$$E(u|X) = 0,$$

where u is the error term and X represents the regressors. This assumption ensures that the expected value of the error term, conditional on the regressors, is zero, meaning there is no systematic relationship between X and u.

Implication of Exogeneity:
When $E(u|X) = 0$ holds, the OLS estimator for β:

$$\hat{\beta} = (X'X)^{-1}X'y$$

is unbiased and consistent, meaning that it correctly estimates the true parameter values as the sample size increases.

Stochastic vs. Non-Stochastic Regressors The assumption that X is non-stochastic (fixed in repeated samples) simplifies OLS estimation, making the classical assumptions hold. However, in many practical cases, X is stochastic, meaning it is determined within the system being studied. If X is correlated with the error term ($E(u|X) \neq 0$), endogeneity arises, leading to biased and inconsistent OLS estimates.

Sources of Endogeneity:
Endogeneity can arise due to several reasons:

- **Omitted Variable Bias:** If an important explanatory variable is left out of the model and is correlated with both X and y, this introduces bias.
- **Measurement Error:** Errors in measuring explanatory variables can lead to correlation between X and u.
- **Simultaneous Causality:** If there is a bidirectional relationship between X and y, then X is endogenous.

Example: Education and Wages:
A classical example of endogeneity arises in estimating the return to education on wages. Consider the simple regression model:

$$WAGE_t = \beta_1 + \beta_2 EDUC_t + u_t.$$

If unobserved factors such as ability, motivation, or family background affect both education and wages, then the exogeneity assumption is violated. For instance, if more able individuals tend to acquire more education, then education ($EDUC_t$) is correlated with the error term (u_t), leading to an upward bias in β_2.

Omission of an Important Variable or Inclusion of an Irrelevant Variable:
Omission of an Important Variable

- **Consequence:** The estimated coefficients on all the other variables will be biased and inconsistent unless the excluded variable is uncorrelated with all the included variables.
- Even if this condition is satisfied, the estimate of the coefficient on the constant term will be biased.
- The standard errors will also be biased.

Inclusion of an Irrelevant Variable

- Coefficient estimates will still be consistent and unbiased, but the estimators will be inefficient.

Measurement Error:
Measurement error arises when the observed values of a variable differ from their true values. This issue is particularly significant in econometric models, as it can lead to biased and inconsistent parameter estimates, contributing to endogeneity problems.

Measurement Error in the Explained Variable:
Measurement error in the dependent variable (denoted as y) is generally less problematic than in the explanatory variables. The primary effects include:

- The disturbance term in the regression model accounts for part of the measurement error.
- The estimated parameters remain unbiased and consistent.
- However, standard errors are inflated due to additional noise, leading to larger confidence intervals.

Measurement Error in the Explanatory Variable:
Measurement error in an independent variable (x_t) is more severe as it introduces correlation between the error term and the explanatory variable, violating the classical assumptions of Ordinary Least Squares (OLS). Suppose we estimate the model:

$$y_t = \beta_1 + \beta_2 x_t + u_t,$$

where x_t is measured with error, such that we observe:

$$\tilde{x}_t = x_t + v_t,$$

where v_t is an independent noise term. Substituting into the model:

$$y_t = \beta_1 + \beta_2(\tilde{x}_t - v_t) + u_t,$$

which simplifies to:

$$y_t = \beta_1 + \beta_2 \tilde{x}_t + u_t - \beta_2 v_t.$$

Since $\tilde{x}_t$ is now correlated with the composite error term ($u_t - \beta_2 v_t$), the OLS estimator for β_2 becomes inconsistent. The key implications are:

- The estimated coefficient β_2 is biased towards zero (attenuation bias).
- The bias depends on the variance of the measurement error relative to the variance of x_t.
- If $\beta_2 > 0$, the bias is negative; if $\beta_2 < 0$, the bias is positive.

Sources of Measurement Error:
Measurement errors arise in various circumstances, including:

- Estimation errors in macroeconomic variables (e.g., GDP, inflation) and financial data from company accounts.

- Use of proxy variables when direct measurement is not feasible (e.g., expected inflation instead of actual expectations).

Consequences of Endogenity:
If the exogeneity assumption is violated, the Ordinary Least Squares (OLS) estimator becomes biased and inconsistent. To address this issue, several methods can be applied:

- **Control for Additional Variables**: Endogeneity may result from omitted variables. Including these variables in the model, when observable and available, can mitigate bias.
- **Panel Data Methods**: Techniques such as Fixed Effects or Random Effects can help control for unobserved heterogeneity that may be causing endogeneity.
- **Instrumental Variables (IV) Estimation**: This method involves finding an instrumental variable that is correlated with the endogenous explanatory variable but uncorrelated with the error term, allowing for consistent estimation using approaches such as Two-Stage Least Squares (2SLS).
- **Heckman Correction (Two-Step Procedure)**: This method is used in cases of sample selection bias, a specific form of endogeneity.
- **Difference-in-Differences (DiD)**: If repeated observations are available before and after a treatment or policy change, this approach can help control for unobserved variables that remain constant over time.
- **Structural Modeling**: By constructing a model based on underlying economic theory, structural approaches can help address endogeneity concerns.

Instrumental Variables:
Instrumental Variables (IV) estimation is a technique used to address endogeneity by finding an instrument that allows for consistent estimation. Suppose we are interested in estimating the causal effect of X on Y in the following model:

$$Y_i = \beta_0 + \beta_1 X_i + u_i,$$

where X_i is endogenous, meaning it is correlated with the error term u_i. To resolve this issue, we introduce an instrumental variable Z_i, which satisfies two key conditions:

- **Relevance**: Z_i must be correlated with X_i, i.e., $Cov(Z, X) \neq 0$.
- **Exogeneity**: Z_i must be uncorrelated with the error term u_i, i.e., $Cov(Z, u) = 0$.

The IV estimation proceeds in two stages (Two-Stage Least Squares, 2SLS):
First Stage: Regress X on the instrument Z:

$$X_i = \pi_0 + \pi_1 Z_i + v_i.$$

Second Stage: Use the predicted values from the first stage, $\hat{X}_i$, to estimate:

$$Y_i = \beta_0 + \beta_1 \hat{X}_i + u_i.$$

Example: Military Service and Earnings:
A well-known application of IV is Angrist's (1990) study on the effect of military service on earnings using the Vietnam Era Draft Lottery as an instrument. The key steps are:

- The endogenous regressor: Military Service (X_i) is likely correlated with unobserved determinants of earnings.
- The instrument: The draft lottery number (Z_i) is random and affects the likelihood of serving but is uncorrelated with individual earnings potential.
- Applying 2SLS:
  1. First stage: Predict military service using the draft lottery number.
  2. Second stage: Estimate the effect of predicted military service on earnings.

This method helps obtain an unbiased estimate of the causal effect of military service on earnings. However, finding good instruments remains a challenge in empirical research.

Sample Selection Bias and the Heckman Correction:
Sample selection bias can be addressed using the *Heckman two-step correction*, which accounts for non-random selection into the sample. This method involves two steps:

1. **Selection Model:** The first step estimates the probability of an observation being included in the sample using a selection equation:

   $$S^* = Z\gamma + \nu$$

   where:

   - S^* is a latent variable representing the selection mechanism,
   - Z is a vector of explanatory variables affecting selection,
   - γ is the coefficient vector,
   - ν is the error term.

The observed selection indicator S is defined as:

$$S = \begin{cases} 1, & \text{if } S^* > 0 \\ 0, & \text{otherwise} \end{cases}$$

The selection model is estimated using a probit regression.

2. **Main Model:** The second step estimates the main equation of interest while including a correction term:

$$Y = X\beta + \lambda IMR + \epsilon$$

where:

- Y is the dependent variable,
- X is the set of independent variables,
- β is the coefficient vector,
- λ is the coefficient for the Inverse Mills Ratio (IMR),
- $IMR = \frac{\phi(Z\gamma)}{\Phi(Z\gamma)}$ is derived from the first-step probit model,
- ϵ is the error term.

The inclusion of IMR accounts for the non-random nature of the sample.

Example: IPO Performance and Selection Bias
Consider the study of IPO (Initial Public Offering) performance, where selection bias arises because firms that choose to go public may be systematically different from those that do not. The Heckman correction can be applied as follows:

- **Step 1 (Selection Model):** The dependent variable is a binary indicator of whether a firm has issued an IPO (1 if IPO, 0 otherwise). Independent variables may include firm size, industry, market conditions, and regulatory environment.
- **Step 2 (Main Model):** The dependent variable is a measure of post-IPO performance (e.g., stock returns, return on assets). Independent variables include firm characteristics, market conditions, financial indicators, and the IMR correction term from step 1.

If the IMR coefficient is significant, it indicates that selection bias was present and that the corrected estimates should be used.

Difference-in-Differences (DiD) Estimation:
Difference-in-Differences (DiD) is a quasi-experimental design used to estimate causal relationships. It isolates the impact of a specific intervention or event (treatment) on an outcome by comparing changes over time between a treatment group and a control group.

Basic Assumption: Parallel Trends:
The key assumption of DiD is the *Parallel Trends Assumption*: In the absence of treatment, the treatment and control groups would have followed parallel paths over time.

DiD Estimator:
The DiD estimator is calculated as:

$$\begin{aligned}\text{DiD Estimator} = &\left(\text{Post-Treatment Mean}_{\text{Treatment Group}} - \text{Pre-Treatment Mean}_{\text{Treatment Group}}\right) \\ &- \left(\text{Post-Treatment Mean}_{\text{Control Group}} - \text{Pre-Treatment Mean}_{\text{Control Group}}\right)\end{aligned}$$

Example: Financial Regulation and Bank Profitability Consider evaluating the impact of a new financial regulation on bank profitability. The study setup includes:

- **Treatment Group:** Banks affected by the new regulation.
- **Control Group:** Banks not affected by the regulation.
- **Pre-Treatment Period:** Years before the regulation.
- **Post-Treatment Period:** Years after the regulation.

The hypothesis is that the new regulation will decrease bank profitability (measured by Return on Assets, ROA). The regression model is:

$$ROA_{it} = \beta_1 + \beta_2 \text{TreatGroup}_i + \beta_3 \text{PostRegulation}_t + \delta(\text{TreatGroup}_i \times \text{PostRegulation}_t) + u_{it}$$

DiD Analysis:

- Calculate the average ROA for both groups, before and after the regulation, as a preliminary comparison.
- Apply the DiD equation to estimate the regulation's impact.

Interpretation:

- A negative δ would support the hypothesis that the regulation decreased profitability.
- A positive δ would suggest an unexpected increase in profitability due to the regulation.

Structural Models:

- **Structural** explicitly incorporates economic theory into the model structure.
  - In contrast to reduced-form modeling considered so far, which focuses on estimating relationships between variables without necessarily grounding them in a theoretical framework.

- Many structural models involve systems of simultaneous equations that represent different aspects of the economic theory (e.g., supply and demand, equilibrium conditions).
- Pros and cons
  - Allows for a deeper understanding of the underlying economic mechanisms and policy analysis.
  - Can be complex and computationally intensive, require strong assumptions and strong theory foundation.
- **Example:** Black-Scholes model

3.12.6 Normality: Assumption 5

Bera-Jarque Test To test whether the normality assumption holds, the Bera-Jarque (BJ) test is commonly used. This test examines whether the skewness and kurtosis of the distribution deviate from those of a normal distribution.

Properties of a Normal Distribution:
A normal distribution has the following characteristics:

- It is symmetric (no skewness).
- It has a kurtosis of 3 (mesokurtic distribution).

The BJ test evaluates whether the coefficients of skewness and excess kurtosis are jointly equal to zero:

$$b_1 = \frac{E[u^3]}{(\sigma^2)^{3/2}}, \quad b_2 = \frac{E[u^4]}{(\sigma^2)^2}$$

where:

- b_1 is the skewness coefficient, which measures the asymmetry of the distribution.
- b_2 is the kurtosis coefficient, which measures the heaviness of the tails.

The BJ test statistic is given by:

$$W = T\left[\frac{b_1^2}{6} + \frac{(b_2 - 3)^2}{24}\right] \sim \chi^2(2)$$

where T is the sample size. Under the null hypothesis of normality, W follows a chi-square distribution with two degrees of freedom.

What if Normality is Rejected?:
If the BJ test indicates non-normality, the following considerations should be made:

- It is often unclear what should be done, as many standard techniques assume normality.
- Non-parametric methods could be used, but they may be complex and require additional assumptions.
- Sometimes, one or two extreme residuals cause the rejection of normality; in such cases, further analysis of outliers is necessary.
- A common alternative is to introduce dummy variables to account for specific outliers.

For example, if we estimate a model of asset returns for the period 1980-1990 and identify a large outlier in October 1987 (Black Monday), we may introduce a dummy variable to mitigate its impact.

3.12.7 Multicollinearity

Multicollinearity is a problem that arises in regression analysis when the explanatory variables are highly correlated with each other. This can lead to issues in estimating coefficients, making it difficult to determine the independent effect of each variable.

Perfect Multicollinearity: Perfect multicollinearity occurs when one explanatory variable is a perfect linear combination of another. In such cases:

- It is impossible to estimate all the coefficients in the regression model.
- For example, if $x_3 = 2x_2$, then the design matrix $(X'X)$ is not of full rank and hence not invertible.

Problems Caused by Near Multicollinearity:
When multicollinearity is present but not perfect, it can still cause significant issues:

- The coefficient of determination (R^2) may be high, but individual regression coefficients will have large standard errors.
- The regression results become highly sensitive to small changes in the model specification.
- Confidence intervals for the estimated coefficients become very wide, making significance tests unreliable.

Measuring Multicollinearity:
The simplest way to detect multicollinearity is by examining the correlation matrix of the explanatory variables. For example:

$$\text{Corr} = \begin{bmatrix} - & 0.2 & 0.8 \\ 0.2 & - & 0.3 \\ 0.8 & 0.3 & - \end{bmatrix}$$

However, if three or more variables exhibit linear dependence (e.g., $x_{2t} + x_{3t} = x_{4t}$), simple pairwise correlations may not be sufficient.
A more robust measure of multicollinearity is the Variance Inflation Factor (VIF), given by:

$$VIF_i = \frac{1}{1 - R_i^2}$$

where R_i^2 is obtained from an auxiliary regression of explanatory variable i on all other independent variables.

Solutions to Multicollinearity: Several approaches exist to address multicollinearity:

- Traditional techniques such as ridge regression or principal components analysis (PCA) can be used, though they introduce additional complexity.
- Some econometricians argue that if the model performs well overall, multicollinearity can simply be ignored.
- Practical remedies include:
  - Dropping one of the collinear variables.
  - Transforming highly correlated variables into ratios.
  - Collecting more data to reduce dependency between variables.
  - Using higher-frequency data if applicable.

Overall, addressing multicollinearity depends on the specific problem at hand and the objectives of the regression analysis.

3.13 Principal Component Analysis: PCA

Principal Component Analysis (PCA) is a statistical technique used for dimensionality reduction while preserving as much variance as possible in the data. It is widely used in finance, machine learning, and statistics to transform a set of correlated variables into a new set of uncorrelated variables called *principal components*.

3.13.1 Mathematical Formulation

Given a dataset with k explanatory variables $X_1, \ldots, X_k$, PCA transforms them into k new variables $p_1, \ldots, p_k$, known as principal components, which are linear combinations of the original variables:

$$p_1 = \alpha_{11}X_1 + \alpha_{12}X_2 + \cdots + \alpha_{1k}X_k \quad (1)$$
$$p_2 = \alpha_{21}X_1 + \alpha_{22}X_2 + \cdots + \alpha_{2k}X_k \quad (2)$$
$$\vdots \quad (3)$$
$$p_k = \alpha_{k1}X_1 + \alpha_{k2}X_2 + \cdots + \alpha_{kk}X_k \quad (4)$$

where α_{ij} are coefficients to be determined, known as factor loadings.

3.13.2 Principal Components as Eigenvalues

PCA can be understood as an eigenvalue decomposition of the covariance matrix $X'X$ (or correlation matrix, if standardization is applied). If the ordered eigenvalues of $X'X$ are denoted as λ_i, then the proportion of total variance explained by principal component i is:

$$\phi_i = \frac{\lambda_i}{\sum_{i=1}^{k} \lambda_i}$$

Typically, only the first r $(0 < r < k)$ principal components are retained to capture most of the variation while discarding the remaining $k - r$ components.

3.13.3 PCA and Yield Curve Interpretation:

This is the term structure of interest rates ("yield curve") is a representation that plots bonds of the same type in terms of their yields over different maturities.

- **PC1 - Level:** The first principal component tends to have equal values across all maturities, so controls the overall level of the curve.
- **PC2 - Slope:** The second component usually has a positive relationship with maturity, which allows this component to determine the slope of the curve and the bulk of the differences in yield between short and long maturities.
- **PC3 - Curvature:** The third and usually last component tends to determine how bowed the curve is, leading to it being referred to as curvature.

3.14 Related Topics: Diagnostics

3.14.1 Parameter Stability:

In regression analysis, we often estimate models of the form:

$$y_t = \beta_1 + \beta_2 x_{2t} + \beta_3 x_{3t} + u_t \tag{5}$$

Implicitly, we assume that the parameters $(\beta_1, \beta_2, \beta_3)$ remain constant over the entire sample period. However, this assumption may not hold in practice due to structural changes or regime shifts in the data. To test this assumption, we employ parameter stability tests.

There are two primary tests for parameter stability:

- **Chow Test** (Analysis of Variance Test)
- **Predictive Failure Tests**

The Chow Test:
The Chow test examines whether the regression coefficients are stable across different time periods. The steps to perform the test are:

1. Split the data into two sub-periods and estimate the regression separately for each sub-period, as well as for the entire period. This results in three regressions.

2. Compute the residual sum of squares (RSS) for each regression:

   - RSS for the full-period regression: RSS
   - RSS for the first sub-period: RSS_1
   - RSS for the second sub-period: RSS_2

3. Formulate the F-test statistic as:

$$F = \frac{RSS - (RSS_1 + RSS_2)}{RSS_1 + RSS_2} \times \frac{T - 2k}{k} \sim F(k, T - 2k) \tag{6}$$

   where T is the total number of observations and k is the number of estimated parameters.

If the test statistic exceeds a critical value from the F-distribution, we reject the null hypothesis of parameter stability, indicating a structural break.

The Predictive Failure Test:
A limitation of the Chow test is that it requires sufficient data in both subsamples, i.e., $T_1 \gg k, T_2 \gg k$. An alternative approach is the predictive failure test, which assesses how well a model estimated in one period predicts observations in another period.
The steps for the predictive failure test are:

1. Estimate the regression using the full sample and obtain the RSS.

2. Estimate the regression using only a "large" sub-period and compute its RSS, denoted RSS_1.

3. Compute the test statistic:

$$F = \frac{RSS - RSS_1}{RSS_1} \times \frac{T_1 - k}{T_2} \tag{7}$$

   where T_2 is the number of observations being predicted.

The test statistic follows an $F(T_2, T_1 - k)$ distribution.

Choosing Sub-Samples for Testing:
To determine appropriate sub-samples for stability testing, we can use:

- Visual inspection: Plot the dependent variable over time and identify structural breaks.

- Historical events: Consider known events such as financial crises or policy changes.
- Predictive failure method: Exclude the last few observations and test their predictability.

These methods help in detecting parameter instability and ensuring robustness in regression models.

3.14.2 Building Econometric Models

Objectives of an Econometric Model:
Our objective is to build a statistically adequate empirical model that:

- Satisfies the assumptions of the Classical Linear Regression Model (CLRM).
- Is logically plausible.
- Has an appropriate theoretical interpretation.
- Has the right "shape," meaning:
  - All signs on coefficients are "correct."
  - All sizes of coefficients are "correct."
- Is capable of explaining the results of all competing models.

The General-to-Specific Approach:
The general-to-specific (GETS) approach consists of the following steps:

1. Form a "large" model with many explanatory variables on the right-hand side. This is known as a Generalised Unrestricted Model (GUM).
2. Ensure that the model satisfies all assumptions of the CLRM.
3. If assumptions are violated, take appropriate corrective actions, such as:
   - Taking logarithms.
   - Adding lagged variables.
   - Introducing dummy variables.
4. These steps must be completed before hypothesis testing.
5. The resulting model may be quite large, including multiple lags and independent variables.

Reparameterizing the Model:
Once an initial model is built, the next stage is to reparameterize it by:

- Removing very insignificant regressors.

- Identifying coefficients that are insignificantly different from each other and combining them.
- Checking at each stage that the assumptions of CLRM are still valid.

Final Model and Applications:
At this stage, we aim to obtain a statistically adequate empirical model that can be used for:

- Testing underlying financial and economic theories.
- Forecasting future values of the dependent variable.
- Formulating policies and decision-making.

4 Time Series Analysis

4.1 Time Series Data

Time series data consists of observations recorded sequentially over time. Understanding the key characteristics of time series data is essential for modeling and forecasting in finance and economics. Below are the fundamental concepts:

- **Stationarity**: A time series is stationary if its statistical properties (mean, variance, autocorrelation) remain constant over time. Many financial models assume stationarity.
- **Nonstationarity**: A nonstationary time series exhibits trends, seasonality, or changing variance. Price levels are typically nonstationary, while returns are often stationary.
- **Autocorrelation**: Measures the correlation between observations at different time lags. Returns often exhibit low autocorrelation, whereas squared returns show significant autocorrelation (volatility clustering).
- **White Noise**: A series of uncorrelated random variables with constant mean and variance. Financial returns are often modeled as white noise.
- **Random Walk**: A nonstationary process where each observation is the sum of the previous value and a random shock. Asset prices are often modeled as random walks.
- **Volatility Clustering**: Large price movements tend to be followed by large movements, and small movements follow small movements. This phenomenon leads to time-varying volatility in financial markets.
- **Fat Tails**: Financial return distributions often exhibit excess kurtosis (leptokurtosis), meaning extreme events occur more frequently than in a normal distribution.

- **Leverage Effect**: Negative returns (price drops) tend to increase volatility more than positive returns of the same magnitude.
- **Mean Reversion**: A tendency for a time series to return to a long-term mean. Common in interest rates and volatility models but not typically in asset prices.
- **Seasonality**: A repeating pattern in the data at regular intervals, common in macroeconomic and business cycle data.

4.2 Time Series Models - Introduction

The fundamental problem in time series modeling is to express the dependent variable Y as a function of the independent variable(s) X. Formally, we write:

$$Y = f(X) + \epsilon, \tag{8}$$

where $f(X)$ represents the underlying functional relationship, and ϵ is an error term accounting for randomness and model imperfections.

A key challenge in this context is determining:

- The appropriate predictors, i.e., the content of X.
- The suitable functional form of f.

An important consideration is the bias-variance trade-off. Given an estimate of f based on a set of predictors X, the prediction is $\hat{Y} = \hat{f}(X)$. Assuming that f and X are fixed while variability arises solely from ϵ, the expected squared error can be decomposed as:

$$\mathbb{E}(Y - \hat{Y})^2 = \mathbb{E}\big[(f(X) + \epsilon - \hat{f}(X))^2\big] \tag{9}$$

$$= \underbrace{(f(X) - \hat{f}(X))^2}_{\text{Reducible}} + \underbrace{\text{Var}(\epsilon)}_{\text{Irreducible}} . \tag{10}$$

This decomposition highlights the trade-off between bias (the error due to incorrect model assumptions) and variance (the sensitivity of the model to variations in training data). An overly flexible model may reduce bias but increase variance, leading to overfitting.
Increased model flexibility can improve in-sample fit but often degrades performance on out-of-sample data. This phenomenon is known as overfitting.

There are numerous approaches to specifying the function f, depending on the characteristics of the time series:

- **Machine Learning**
  - Regularized regressions
  - Tree-based methods

  - Deep learning

- **Multivariate Models**
  - Linear and non-linear regressions
  - Mixed Data Sampling (MIDAS)

- **Naive Methods**
  - Random walk

- **Univariate Time Series Models**
  - Autoregressive (AR)
  - Moving Average (MA)
  - ARMA, ARIMA, ARIMAX, SARIMA

Each of these approaches has strengths and weaknesses, and the choice depends on the nature of the data and the forecasting objective.

4.2.1 Short Introduction: Univariate Time Series Models

A **time series** is a sequence of observations recorded over time, denoted as $\{y_1, y_2, \ldots, y_T\}$. The goal of time series modeling is to estimate the conditional expectation $E[y_t|F_{t-1}]$, where F_{t-1} represents the past history of the series.

The ARIMA Class of Models:
One of the most widely used frameworks for modeling univariate time series is the **ARIMA** (AutoRegressive Integrated Moving Average) class of models. These models aim to capture the dependence structure in a time series using past values and error terms. The key models in this class include:

- **AutoRegressive (AR) model** – relies on past values of the series.
- **Moving Average (MA) model** – incorporates past error terms.
- **AutoRegressive Moving Average (ARMA) model** – combines both AR and MA components.
- **AutoRegressive Integrated Moving Average (ARIMA) model** – extends ARMA by including differencing to handle non-stationarity.

Applications and Use Cases:
The ARIMA class of models is particularly useful in financial time series analysis, where variables often exhibit strong temporal dependencies. These models:

- Use only past values of the variable and possibly past error terms.
- Capture key empirical properties of observed data.

- Are beneficial when explanatory variables are unavailable or not measurable.
- Can handle different data frequencies and support out-of-sample forecasting.

Due to their versatility, ARIMA models are widely used in financial markets, economic forecasting, and risk management.

4.3 Seasonality

Seasonality refers to patterns in time series data that repeat at regular intervals due to systematic influences such as economic cycles, weather conditions, or recurring business activities. These patterns can cause fluctuations that obscure underlying trends and make forecasting more challenging. Properly handling seasonality is crucial for improving predictive models and making better-informed financial decisions.

4.3.1 Methods for Dealing with Seasonality

There are several approaches to handling seasonality in time series data, broadly categorized into preprocessing techniques, explicit seasonality modeling, feature engineering, resampling, and machine learning methods.

Preprocessing Techniques:

- **Seasonal Differencing:** Subtracting the value from the same season in the previous cycle to remove seasonality.
- **Log-Transforms:** Applying logarithms to stabilize variance in data.
- **Moving Averages:** Smoothing data by averaging over a fixed time window to remove short-term fluctuations.
- **Decomposition:** Splitting the time series into trend, seasonal, and residual components for separate analysis.

Explicit Seasonality Modeling:

- **SARIMA (Seasonal Autoregressive Integrated Moving Average):** Extends ARIMA to capture seasonal components.
- **Holt-Winters Method:** A smoothing technique that accounts for trend and seasonality.
- **TBATS Model:** Incorporates trigonometric terms, Box-Cox transformations, ARMA errors, and trend components to handle complex seasonal patterns.

Feature Engineering:

- **Dummy Variables:** Encoding categorical seasonal effects (e.g., months, quarters).
- **Fourier Terms:** Using sine and cosine terms to model periodic fluctuations.
- **Lags:** Including previous observations as predictors.
- **External Regressors:** Incorporating exogenous variables that influence seasonality.

Resampling and Aggregation:

- **Temporal Aggregation:** Converting high-frequency data (e.g., daily) to lower frequency (e.g., weekly or monthly) to reduce short-term seasonal noise.

Machine Learning Models:

- **Recurrent Neural Networks (RNNs):** Capable of capturing complex sequential dependencies.
- **Tree-Based Methods:** Random forests and gradient boosting models can capture non-linear seasonal patterns.

We look into two methods in more detail; STL decomposition and seasonal differencing.

4.3.2 STL Decomposition:

A time series decomposition can be used to measure the strength of trend and seasonality in a time series. STL (Seasonal-Trend decomposition using LOESS) is a widely used decomposition method that allows flexible estimation of trend and seasonal components. The decomposition is expressed as:

$$y_t = T_t + S_t + R_t, \tag{11}$$

where:

- T_t is the smoothed trend component,
- S_t is the seasonal component,
- R_t is the remainder (or residual) component.

The STL decomposition method uses locally weighted regression (LOESS) to estimate the trend and seasonal components iteratively, making it robust to outliers and allowing it to handle nonstationary seasonality. This method is advantageous because it enables control over the smoothness of the trend and seasonal components separately.

For strongly trended data, the seasonally adjusted data ($y_t - S_t$) should exhibit significantly more variation than the remainder component. Conversely, if the data has weak trend strength, the remainder component may dominate the variation.

STL decomposition is particularly useful for:

- Handling complex seasonal patterns that may change over time,
- Being robust to missing values and outliers,
- Allowing for different seasonal lengths (e.g., weekly, monthly, or quarterly trends).

4.3.3 Seasonal Differencing

Seasonal differencing is a technique used in time series analysis to remove seasonality from a dataset. It helps make the series stationary, which is a crucial assumption for many time series models.

What is Seasonal Differencing?:

- A method to eliminate seasonality in a time series.
- Helps in making the series stationary for modeling purposes.
- Involves subtracting a value from its previous seasonal lag.

Mathematical Definition:
Seasonal differencing transforms a time series Y_t by subtracting its value at a seasonal lag s, leading to a new series Y_t':

$$Y_t' = Y_t - Y_{t-s} \tag{12}$$

where:

- Y_t is the original time series.
- s is the seasonal period (e.g., $s = 12$ for monthly data with yearly seasonality).
- Y_t' is the seasonally differenced series.

Interpretation and Importance:
Seasonal differencing is crucial for transforming non-stationary time series into a stationary format, which is required for various modeling techniques such as ARIMA. We use seasonal differencing to;

- Removes repeating seasonal patterns, making trends clearer.
- Helps in identifying trends and ensuring stationarity.
- Prepares data for advanced time series models like SARIMA and ARIMA.

Applying seasonal differencing correctly ensures that the underlying patterns in the data are preserved while eliminating seasonal fluctuations, thereby improving forecasting accuracy.

4.4 Stationarity

4.4.1 Definitions

Stationarity is a fundamental concept in time series analysis that ensures the statistical properties of a process remain unchanged over time. There are two key types of stationarity: strict stationarity and weak (or covariance) stationarity.

Strict Stationarity:
A stochastic process $\{y_t\}$ is said to be **strictly stationary** if its joint distribution remains unchanged when shifted in time. Formally, for any set of time indices $t_1, t_2, \ldots, t_n$ and any time shift m, we have:

$$P\{y_{t_1} \leq b_1, \ldots, y_{t_n} \leq b_n\} = P\{y_{t_1+m} \leq b_1, \ldots, y_{t_n+m} \leq b_n\}. \tag{13}$$

This implies that the probability distribution of any subset of the time series does not change over time.

Weak Stationarity (Covariance Stationarity):
A stochastic process $\{y_t\}$ is said to be **weakly stationary** (or **covariance stationary**) if it satisfies the following three conditions:

1. Constant mean: $E(y_t) = \mu, \quad \forall t.$
2. Constant variance: $E(y_t - \mu)^2 = \sigma^2 < \infty, \quad \forall t.$
3. Autocovariance depends only on lag: $E(y_{t_1} - \mu)(y_{t_2} - \mu) = \gamma_{t_2 - t_1}, \quad \forall t_1, t_2.$

Thus, in a weakly stationary process, the mean and variance are time-invariant, and the covariance between two time points depends only on the time difference (lag) between them.

Autocovariance Function:
The autocovariance function describes the relationship between different points in a weakly stationary time series. It is defined as:

$$E(y_t - E(y_t))(y_{t-s} - E(y_{t-s})) = \gamma_s, \quad s = 0, 1, 2, \ldots \tag{14}$$

where γ_s represents the covariance at lag s.

Autocorrelation Function:
The autocorrelation function (ACF) normalizes the autocovariance function to remove the dependency on units. It is given by:

$$\tau_s = \frac{\gamma_s}{\gamma_0}, \quad s = 0, 1, 2, \ldots \tag{15}$$

where γ_0 is the variance of the process. The autocorrelation function measures the linear relationship between observations at different time points, standardized to be between -1 and 1.

4.4.2 The "Random" Walk:

Before we can define the random walk we must define white noise.

White Noise:
A white noise process is a stochastic process with (virtually) no structure. Formally, a time series $\{y_t\}$ is said to be a white noise process if it satisfies the following properties:

- The mean of the process is constant:

$$E(y_t) = \mu. \tag{16}$$

- The variance is constant:

$$\text{Var}(y_t) = \sigma^2. \tag{17}$$

- The autocovariance function satisfies:

$$\gamma_{t-r} = \begin{cases} \sigma^2, & \text{if } t = r, \\ 0, & \text{otherwise.} \end{cases} \tag{18}$$

This implies that the random variables y_t are independent and identically distributed (i.i.d.). In other words, there is no dependence between observations at different time periods. White noise processes are fundamental in time series analysis and serve as building blocks for more complex models.
Additionally, white noise processes are examples of **stationary** processes, as they exhibit constant mean and variance over time, and their autocovariance depends only on the lag between observations.

Gaussian White Noise:
If the random variables y_t follow a normal distribution, i.e.,

$$y_t \sim \mathcal{N}(\mu, \sigma^2), \tag{19}$$

then the process is called *Gaussian white noise*. Gaussian white noise is commonly used in financial modeling and econometrics as an assumption for residuals in regression models and stochastic processes.

Significance Testing for White Noise:

If we assume that y_t are i.i.d. with a standard normal distribution, and T is the sample size, then the sample autocorrelation coefficient $\hat{\rho}_s$ follows approximately:

$$\hat{\rho}_s \sim \mathcal{N}\left(0, \frac{1}{T}\right). \tag{20}$$

Using this result, we can perform significance tests for the autocorrelation coefficients by constructing confidence intervals. For example, a 95% confidence interval is given by:

$$\pm 1.96 \times \frac{1}{\sqrt{T}}. \tag{21}$$

If the sample autocorrelation coefficient, $\hat{\rho}_s$, falls outside this interval for any value of s, we reject the null hypothesis that the true value of the autocorrelation coefficient at lag s is zero. This suggests that the time series exhibits some form of autocorrelation and is not purely white noise.

Random Walk:
A random walk is a stochastic process where the current value of a series is equal to the previous value plus a random shock. Formally, a random walk process is defined as:

$$y_t = y_{t-1} + u_t, \tag{22}$$

where u_t is white noise, meaning that it is an independent and identically distributed (i.i.d.) random variable with mean zero and constant variance.

Stationarity of a Random Walk:
Random walks are prominent examples of *non-stationary* processes. A process is stationary if its statistical properties, such as mean and variance, do not change over time. However, in a random walk:

- The mean of y_t depends on its initial value y_0, making it time-dependent.
- The variance of y_t increases over time, as the cumulative sum of u_t leads to an unbounded variance.
- The autocorrelation function (ACF) does not decay quickly, as opposed to stationary processes where correlations diminish over time.

Autocorrelation Function of a Random Walk:
The autocorrelation function (ACF) of a random walk demonstrates the fundamental difference between white noise and a random walk.

- For white noise, the autocorrelations are close to zero at all lags, indicating no dependence between observations.
- For a random walk, the autocorrelation function exhibits slow decay, implying strong dependence over time.

Illustration:

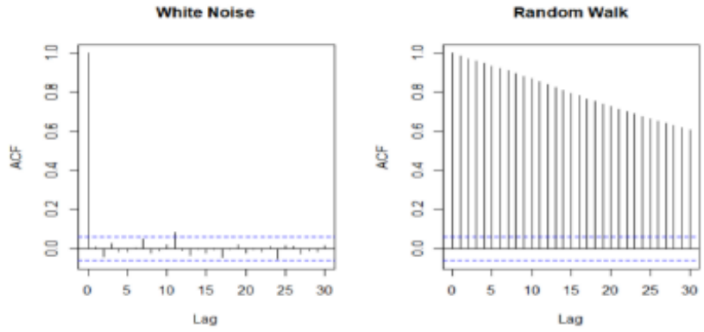


Figure 3: Autocorrelation function of a simulated random walk

This fundamental property of random walks is why they are often used to model financial time series, such as stock prices, where past values strongly influence future values.

4.4.3 Unit Roots and Non-Stationarity

Definition of Non-Stationarity:
A time series is non-stationary if its statistical properties, such as mean and variance, change over time. A series is said to be integrated of order d, denoted as $I(d)$, if it must be differenced d times to become stationary. That is,

- If $y_t \sim I(d)$, then $\Delta^d y_t \sim I(0)$.
- An $I(0)$ series is stationary, meaning its properties do not change over time.

Economic and financial time series often contain a single unit root ($I(1)$), while some macroeconomic indicators, such as price indices, may exhibit $I(2)$ behavior.

Types of Non-Stationarity:
Non-stationarity can arise due to different mechanisms. Two commonly studied models are:

- **Random Walk with Drift**: A stochastic process defined as

$$y_t = \mu + y_{t-1} + u_t, \quad u_t \sim \text{iid}. \tag{23}$$

  This model exhibits stochastic non-stationarity, where shocks persist indefinitely, and the variance increases over time.

- **Deterministic Trend Process**: A series with a deterministic time trend:

$$y_t = \alpha + \beta t + u_t. \tag{24}$$

  This series is trend-stationary, meaning stationarity can be achieved by removing the deterministic trend.

Dealing with Non-Stationarity:
To make a non-stationary series stationary, different transformations must be applied depending on the nature of the process:

- **Random walk with drift**: Apply first differencing ($\Delta y_t = y_t - y_{t-1}$) to remove the stochastic trend.
- **Trend-stationary process**: Detrend the series by regressing on time and extracting residuals.

Choosing the incorrect transformation can lead to distortions. For example, differencing a trend-stationary series introduces a moving average structure in the errors, whereas detrending a unit root process does not remove its non-stationarity.

Impact of Shocks on Stationary and Non-Stationary Series:
The behavior of shocks in a time series depends on the autoregressive parameter ϕ in the model:

$$y_t = \phi y_{t-1} + u_t. \tag{25}$$

1. If $|\phi| < 1$, the series is stationary, meaning shocks decay over time as $\phi^T \to 0$.
2. If $\phi = 1$, the series follows a unit root process, meaning shocks accumulate and persist indefinitely:
$$y_t = y_0 + \sum_{t=0}^{\infty} u_t. \tag{26}$$
3. If $\phi > 1$, the series exhibits explosive behavior, where shocks grow over time instead of dissipating.

Deterministic vs. Stochastic Non-Stationarity:
Non-stationarity can be classified as deterministic or stochastic:

- **Deterministic Non-Stationarity**: Driven by a predictable time trend. The trend can be removed through detrending techniques.
- **Stochastic Non-Stationarity**: Arises due to the presence of a unit root. Shocks have permanent effects and must be removed through differencing.

In economic and financial time series, stochastic non-stationarity is more prevalent, making it crucial to conduct unit root tests, such as the Augmented Dickey-Fuller (ADF) test, to distinguish between the two types. Applying the appropriate transformation ensures valid statistical inference and modeling.

4.5 Statistical Tests for Time Series Data

We have two main reasons for applying statistical tests to time series data:

1. Investigating the statistical characteristics of a time series

2. Model residual diagnostics

Note that some of the statistical tests already have been mentioned in the section about the Gauss-Markov Assumptions. Thus, we will not mention all again, but only the ones most common for time series.

4.5.1 Normality

Jarque-Bera *Objective:*
Tests whether a random vector comes from a normal distribution with an unknown mean and variance.

Test statistic:

$$JB = \frac{n}{6}\left(s^2 + \frac{(k-3)^2}{4}\right),$$

where n is the sample size, s is the sample skewness, k is the sample kurtosis. For large sample sizes, the test statistic has a chi-square distribution with two degrees of freedom.

Test hypotheses:

$$H_0: \text{ Data comes from a normal distribution}$$

$$H_A: \text{ Data does not come from a normal distribution}$$

4.5.2 Autocorrelation

See other section for Durbin-Watson.

Box-Pierce:

Test objective:
Test the joint hypothesis that all m of the τ_k correlation coefficients are simultaneously equal to zero.

Test statistic:

$$Q = T\sum_{k=1}^{m}\hat{\tau}_k^2 \sim \chi_m^2$$

where T = sample size, m = maximum lag length.

Test hypotheses:

$$H_0: \text{ A series of observations are i.i.d } \rightarrow \text{ no serial correlation.}$$

$$H_A: \text{ Observations are serially correlated}$$

Ljung-Box:

Test objective:
The Ljung-Box Q-test is a "portmanteau" test that assesses the null hypothesis that a series of residuals exhibits no autocorrelation for a fixed number of lags L.

Test statistic:

$$Q^* = T(T+2)\sum_{k=1}^{m}\frac{\hat{\tau}_k^2}{T-k} \sim \chi_m^2$$

where T = sample size, m = maximum lag length.

Test hypotheses:

$$H_0: \text{ All autocorrelation coefficients } \tau_k, k = 1, \cdots, L \text{ are zero.}$$
$$H_A: \text{ Some autocorrelation coefficient } \tau_k, k = 1, \cdots, L \text{ is nonzero.}$$

4.5.3 Heteroscedacity

Engle's ARCH test:

Test objective:

The ARCH test investigates the presence of conditional heteroscedasticity in residuals.

Test statistic:
The ARCH(L) model has the following form:

$$r_t^2 = \alpha_0 + \alpha_1 r_{t-1}^2 + \cdots + \alpha_L r_{t-L}^2 + e_t,$$

where there is at least one $\alpha_j \neq 0, j = 0, \ldots, L$. The test statistic is the Lagrange multiplier statistic $TR^2 \sim \chi_L$, where

- T is the sample size
- R^2 is the coefficient of determination from fitting the ARCH(L) model for a number of lags (L) via regression.

Test hypotheses:

$$H_0: \text{ a series of residuals } r_t \text{ exhibits no conditional heteroscedasticity.}$$
$$H_A: \text{ an ARCH}(L) \text{ model describes the series.}$$

4.5.4 Unit Roots / Stationarity

Dickey-Fuller (DF) Test:
The DF test is designed to test for a unit root in a simple autoregressive model:

$$y_t = \phi y_{t-1} + u_t \quad \Leftrightarrow \quad \Delta y_t = \psi y_{t-1} + u_t, \tag{27}$$

where $\Delta y_t = y_t - y_{t-1}$ is the first difference of the series, and u_t is a white noise error term. The test statistic is computed as:

$$\text{test statistic} = \frac{\hat{\psi}}{SE(\hat{\psi})}. \tag{28}$$

The null and alternative hypotheses are:

$$\begin{aligned} H_0: &\quad \phi = 1 \Leftrightarrow \psi = 0 \quad \Rightarrow \quad \text{non-stationary} \\ H_A: &\quad \phi < 1 \Leftrightarrow \psi \neq 0 \quad \Rightarrow \quad \text{stationary.} \end{aligned}$$

If the null hypothesis is rejected, the time series is considered stationary.

Augmented Dickey-Fuller (ADF) Test:
The DF test assumes that the error term u_t is white noise. However, in practice, autocorrelation may be present. The ADF test addresses this issue by including lagged differences of the dependent variable:

$$\Delta y_t = \psi y_{t-1} + \sum_{i=1}^{p} \alpha_i \Delta y_{t-i} + u_t. \tag{29}$$

The null and alternative hypotheses remain the same as in the DF test. The inclusion of lagged differences ensures that the residuals are white noise, making the test more robust to autocorrelation.

Phillips-Perron (PP) Test:
The Phillips-Perron test is an alternative to the ADF test that corrects for autocorrelation and heteroskedasticity in the error term without adding lagged differences. The model is given by:

$$y_t = c + \delta t + \phi y_{t-1} + \varepsilon_t, \tag{30}$$

where c is a drift term, δ represents a deterministic trend, and ε_t is a mean-zero innovation process. The PP test applies a non-parametric correction to the standard Dickey-Fuller test statistic, making it useful when the residuals exhibit autocorrelation.

Criticism and Limitations:
While these tests are widely used, they have some limitations:

- The power of the tests is low if the process is stationary but close to the non-stationary boundary (e.g., $\phi = 1$ versus $\phi = 0.95$), especially for small sample sizes.

- They do not handle structural breaks well. If a structural break is present, the null hypothesis (H_0) may not be rejected, even if the series is stationary after accounting for the break.

Testing for Higher Order of Integration:
If H_0 is rejected, we conclude that the time series does not contain a unit root. However, if it is not rejected, we may need to test for a higher order of integration:

$$H_0 : y_t \sim I(2) \quad \text{vs.} \quad H_1 : y_t \sim I(1). \tag{31}$$

Heteroskedasticity and Autocorrelation Consistent (HAC) Standard Errors:
When autocorrelation and/or heteroskedasticity are present in the error terms, standard errors may be inconsistent. HAC estimators provide robust standard errors, commonly using:

- Newey-West estimator
- Hansen-Hodrick estimator

These estimators are particularly useful in financial time series analysis.

4.5.5 Newey-West Estimator

The Newey-West estimator is a widely used method for computing robust standard errors in the presence of heteroskedasticity and autocorrelation. It adjusts the covariance matrix to ensure consistency of estimated standard errors.

The estimator is computed as:

$$\hat{\Omega} = \sum_{j=-q}^{q} w_j \hat{\Gamma}_j, \tag{32}$$

where $\hat{\Gamma}_j$ is the autocovariance at lag j, and w_j is a weight function that decreases with increasing j.

The key features of the Newey-West estimator are:

- It remains consistent even when errors exhibit autocorrelation and heteroskedasticity.
- It uses a Bartlett kernel to assign decreasing weights to higher-order lags.
- The choice of lag truncation parameter q affects efficiency; common choices include $q = 4(T/100)^{2/9}$ for sample size T.

4.6 Moving Average (MA) Processes

A *Moving Average (MA)* process of order q, denoted MA(q), is a fundamental time series model that represents a linear combination of *past error terms* (white noise). MA processes capture *short-term dependencies* in the data.

4.6.1 Definition

Let $\{u_t\}$ be a sequence of i.i.d. random variables with

$$E(u_t) = 0, \quad \text{Var}(u_t) = \sigma^2.$$

A MA(q) process is defined as:

$$y_t = \mu + u_t + \theta_1 u_{t-1} + \theta_2 u_{t-2} + \cdots + \theta_q u_{t-q}, \tag{33}$$

or equivalently,

$$y_t = \mu + \sum_{i=1}^{q} \theta_i \, u_{t-i} + u_t.$$

Here, μ is the mean of the process, θ_i are the MA coefficients, and u_t represents the white-noise error term.

4.6.2 Lag Operator Representation for MA

Define the *lag operator* L by $L^i y_t = y_{t-i}$. For an MA(q) process, we can write

$$y_t = \mu + \theta(L) u_t,$$

where

$$\theta(L) = 1 + \theta_1 L + \theta_2 L^2 + \cdots + \theta_q L^q.$$

Hence,

$$y_t = \mu + \theta(L) u_t.$$

This representation often simplifies derivations of stationarity or invertibility conditions and clarifies the structure of MA processes.

4.6.3 Properties of the MA(q) Process

- **Constant Mean:**

$$E(y_t) = \mu.$$

- **Constant Variance:**

$$\text{Var}(y_t) = \gamma_0 = \big(1 + \theta_1^2 + \theta_2^2 + \cdots + \theta_q^2\big)\sigma^2.$$

- **Autocovariance:** The autocovariance function (ACF) of an MA(q) process is nonzero only for lags up to q:

$$\gamma_s = \begin{cases} \Big(\theta_s + \sum_{j=1}^{q-s} \theta_{s+j}\theta_j\Big)\sigma^2, & \text{for } s = 1, \ldots, q, \\ 0, & \text{for } s > q. \end{cases}$$

  This makes MA processes particularly useful in modeling short-term dependencies.

4.6.4 Invertibility Condition

For the MA(q) process to be *invertible* (i.e., to admit a convergent infinite autoregressive representation), the roots of the characteristic equation

$$1 + \theta_1 z + \theta_2 z^2 + \cdots + \theta_q z^q = 0$$

must lie *outside* the unit circle in the complex plane.

4.6.5 Representation of MA as an Infinite AR Process

When the MA(q) process is *invertible*, it can be expressed as an infinite-order AR process. This is useful if we wish to analyze the persistence of shocks or estimate models using only past y-values rather than unobserved error terms.

Example: MA(1) as AR(∞). Consider the MA(1) process:

$$y_t = u_t + \theta_1 u_{t-1}.$$

If the process is invertible, we can solve for u_t in terms of past y-values:

$$u_t = y_t - \theta_1 u_{t-1}.$$

Substituting u_{t-1} from its own definition and continuing recursively, we obtain:

$$u_t = y_t - \theta_1 y_{t-1} + \theta_1^2 y_{t-2} - \theta_1^3 y_{t-3} + \dots,$$

which yields an infinite-order AR representation:

$$y_t = \theta_1 y_{t-1} - \theta_1^2 y_{t-2} + \theta_1^3 y_{t-3} - \cdots + u_t.$$

4.7 Autoregressive (AR) Processes

An *Autoregressive (AR)* process of order p, denoted AR(p), expresses the current value of a time series as a linear combination of its *past values* and a white noise term. AR models can capture potentially *longer-term* dependencies than finite-order MA models.

4.7.1 Definition

An AR(p) process is given by:

$$y_t = \mu + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + u_t, \tag{34}$$

or equivalently,

$$y_t = \mu + \sum_{i=1}^{p} \phi_i y_{t-i} + u_t,$$

where μ is the mean, ϕ_i are the AR coefficients, and u_t is a white-noise process.

4.7.2 Lag Operator Representation

Define the lag operator L by $L^i y_t = y_{t-i}$. The AR(p) model can then be written as

$$y_t = \mu + \sum_{i=1}^{p} \phi_i L^i y_t + u_t,$$

or equivalently,

$$\phi(L)\, y_t = \mu + u_t,$$

where

$$\phi(L) = 1 - \phi_1 L - \phi_2 L^2 - \cdots - \phi_p L^p.$$

4.7.3 Stationarity Condition

For the AR(p) process to be *stationary*, the roots of the characteristic polynomial

$$1 - \phi_1 z - \phi_2 z^2 - \cdots - \phi_p z^p = 0$$

must lie *outside* the unit circle in the complex plane. A stationary AR(p) admits an MA(∞) representation, consistent with the general idea that any stationary process can be expressed as an infinite MA (see below).

4.7.4 Properties of the AR(p) Process

Key properties of an AR(p) process, assuming stationarity and $E(u_t) = 0$, include:

- **Mean:**
$$E(y_t) \;=\; \frac{\mu}{1 - \sum_{i=1}^{p} \phi_i}.$$

- **Variance:** The unconditional variance can be derived by noting that

$$y_t - E(y_t) \;=\; \sum_{i=1}^{p} \phi_i \big(y_{t-i} - E(y_t)\big) + u_t,$$

  which, after some algebra involving the Yule-Walker equations, yields

$$\mathrm{Var}(y_t) = \frac{\sigma^2}{1 - \sum_{i=1}^{p} \phi_i z_i(\dots)},$$

  where the exact closed-form depends on $\{\phi_i\}$. In practice, one often uses estimates of the autocovariance function or maximum likelihood to evaluate it numerically.

4.7.5 Wold's Decomposition Theorem and AR as MA(∞)

Wold's Decomposition Theorem states that every purely non-deterministic, stationary time series can be represented as an infinite moving average of white noise. In the case of an AR(p) process, it specifically implies that

$$\phi(L)y_t = u_t \quad \Longrightarrow \quad y_t = \phi(L)^{-1}u_t = \psi(L)\, u_t,$$

where

$$\psi(L) = \phi(L)^{-1} = \left(1 - \phi_1 L - \phi_2 L^2 - \cdots - \phi_p L^p\right)^{-1}.$$

Thus,

$$y_t = \sum_{j=0}^{\infty} \psi_j\, u_{t-j},$$

for some coefficients $\{\psi_j\}$.

Example: AR(1). For an AR(1):

$$y_t = \phi_1 y_{t-1} + u_t,$$

the inverse operator $\phi(L)^{-1}$ expands into a geometric series (if $|\phi_1| < 1$), giving

$$y_t = u_t + \phi_1 u_{t-1} + \phi_1^2 u_{t-2} + \ldots,$$

i.e., an MA(∞) representation.

4.8 Autoregressive Moving Average (ARMA) Processes

An *Autoregressive Moving Average (ARMA)* process combines both AR and MA dynamics to capture *short-term* and *longer-term* dependencies in a time series.

4.8.1 Definition

An ARMA(p, q) model is given by:

$$\phi(L)\, y_t = \mu + \theta(L)\, u_t,$$

where

$$\phi(L) = 1 - \phi_1 L - \phi_2 L^2 - \cdots - \phi_p L^p, \quad \theta(L) = 1 + \theta_1 L + \theta_2 L^2 + \cdots + \theta_q L^q.$$

Expanding,

$$y_t = \mu + \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p} + \theta_1 u_{t-1} + \cdots + \theta_q u_{t-q} + u_t,$$

with $\{u_t\}$ white noise satisfying

$$E(u_t) = 0, \quad \mathrm{Var}(u_t) = \sigma^2, \quad E(u_t u_s) = 0 \quad (t \neq s).$$

4.8.2 Stationarity and Invertibility

- **Stationarity:** Requires that the roots of $\phi(z) = 0$ lie *outside* the unit circle.
- **Invertibility:** Requires that the roots of $\theta(z) = 0$ also lie *outside* the unit circle.

4.8.3 Mean of an ARMA Process

If the AR part is stationary, the mean of y_t is

$$E(y_t) = \frac{\mu}{1 - \sum_{i=1}^{p} \phi_i}.$$

4.8.4 Parameter Estimation for ARMA Models

Estimating Pure AR Models using Yule-Walker Equations For pure autoregressive (AR) models, parameter estimation can be done using the *Yule-Walker equations*, which relate the theoretical autocovariances γ_k to the AR coefficients. These equations are derived by multiplying the AR process by lagged values and taking expectations.

For an AR(p) process:

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + u_t, \tag{35}$$

the Yule-Walker system of equations is given by:

$$\begin{bmatrix} \gamma_0 & \gamma_1 & \gamma_2 & \cdots & \gamma_{p-1} \\ \gamma_1 & \gamma_0 & \gamma_1 & \cdots & \gamma_{p-2} \\ \gamma_2 & \gamma_1 & \gamma_0 & \cdots & \gamma_{p-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \gamma_{p-1} & \gamma_{p-2} & \gamma_{p-3} & \cdots & \gamma_0 \end{bmatrix} \begin{bmatrix} \phi_1 \\ \phi_2 \\ \\ \vdots \\ \phi_p \end{bmatrix} = \begin{bmatrix} \gamma_1 \\ \gamma_2 \\ \\ \vdots \\ \gamma_p \end{bmatrix}. \tag{36}$$

To estimate the AR parameters in practice:

1. Compute the *sample autocovariances*:

$$\widehat{\gamma}_k = \frac{1}{T} \sum_{t=k+1}^{T} (y_t - \bar{y})(y_{t-k} - \bar{y}). \tag{37}$$

2. Solve the Yule-Walker system using linear algebra techniques (e.g., matrix inversion).

3. Estimate the variance of the white noise term:

$$\widehat{\sigma}_u^2 = \widehat{\gamma}_0 - \sum_{j=1}^{p} \widehat{\phi}_j \widehat{\gamma}_j. \tag{38}$$

For an AR(1) model, the equation simplifies to:

$$\phi_1 = \frac{\gamma_1}{\gamma_0} = \rho_1, \tag{39}$$

where ρ_1 is the sample autocorrelation at lag 1.

Estimating Pure MA Models For pure moving average (MA) models, we cannot use Yule-Walker equations because MA processes are defined in terms of past white noise values, which are unobservable. Instead, we estimate the parameters by solving equations based on the autocorrelation function (ACF).

For an MA(q) process:

$$y_t = u_t + \theta_1 u_{t-1} + \theta_2 u_{t-2} + \cdots + \theta_q u_{t-q}, \tag{40}$$

the autocorrelation function truncates after lag q, meaning:

$$\rho_k = 0, \quad \forall k > q. \tag{41}$$

To estimate the MA parameters:

1. Compute the sample autocorrelation function $\hat{\rho}_k$.
2. Solve the moment equations relating θ_j to ρ_k.

For an MA(1) model:

$$\rho_1 = \frac{\theta_1}{1 + \theta_1^2}. \tag{42}$$

Solving this quadratic equation gives:

$$\theta_1 = \frac{-1 \pm \sqrt{1 + 4\rho_1^2}}{2\rho_1}. \tag{43}$$

For an MA(2) model, we solve the system:

$$\rho_1 = \frac{\theta_1 + \theta_1\theta_2}{1 + \theta_1^2 + \theta_2^2}, \tag{44}$$

$$\rho_2 = \frac{\theta_2}{1 + \theta_1^2 + \theta_2^2}. \tag{45}$$

Since MA models are not recursive like AR models, parameter estimation often requires solving nonlinear equations, and numerical methods may be necessary for higher-order MA models.

4.8.5 Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF)

The autocorrelation function (ACF) of an ARMA process displays mixed behavior derived from the AR and MA parts. Beyond lag q, the ACF will resemble that of an AR(p) model.

The Partial Autocorrelation Function (PACF) measures the correlation between y_t and y_{t-k} after controlling for intermediate lags. For an AR(p) process, the PACF cuts off after lag p, while for an MA(q) process, it decays geometrically.

4.8.6 Summary of ACF and PACF for AR and MA Processes

- An **AR(p)** process:
  - Has a geometrically decaying ACF.
  - Has PACF spikes up to order p and then cuts off.
- An **MA(q)** process:
  - Has an ACF that spikes up to order q and then cuts off.
  - Has a geometrically decaying PACF.

4.8.7 Building ARMA Models: The Box-Jenkins Approach

The Box-Jenkins approach provides a systematic methodology for identifying, estimating, and diagnosing ARMA models. It consists of three key steps: Identification, Estimation, and Model Diagnostic Checking.

Step 1: Identification:
The first step in building an ARMA model is determining the appropriate order of the autoregressive (AR) and moving average (MA) components.

- The order is determined using graphical procedures such as the Autocorrelation Function (ACF) and the Partial Autocorrelation Function (PACF).
- These plots help distinguish between AR, MA, and ARMA models by examining their characteristic cutoff and decay properties.
- Modern statistical techniques, such as information criteria (AIC, BIC), are often used for more robust order selection.

Step 2: Estimation:
Once the appropriate model order is identified, the parameters are estimated.

- Parameters of the ARMA model can be estimated using methods such as:
  - **Least Squares Estimation** – Used in simpler cases.

- **Maximum Likelihood Estimation (MLE)** – Provides more efficient parameter estimates under normality assumptions.

- Estimation ensures that the model fits the observed time series data optimally.

Step 3: Model Diagnostic Checking:
After estimation, the adequacy of the model must be validated.

- Box and Jenkins suggest two key diagnostic methods:
  - **Deliberate Overfitting** – Testing the sensitivity of parameter estimates by slightly increasing model order.
  - **Residual Diagnostics** – Examining residuals to check for autocorrelation and normality using tools such as:
    - Ljung-Box test for independence.
    - Q-Q plots and histograms for normality assessment.

- A well-specified ARMA model should have uncorrelated residuals with constant variance.

The Box-Jenkins methodology provides a structured approach for fitting ARMA models, ensuring accurate modeling and forecasting of time series data.

4.8.8 Information Criteria for Model Selection: More specifics!

When selecting an appropriate ARMA model, it is crucial to balance model complexity and goodness of fit. Information criteria (IC) provide a systematic approach to this problem by penalizing excessive parameterization.

Identification Using Information Criteria:
Traditionally, model identification relied on graphical procedures such as the ACF and PACF. However, a more robust approach involves using information criteria:

- The variance of estimators is inversely proportional to the number of degrees of freedom.

- Overly complex models may overfit to data-specific noise.

- IC balances goodness of fit with a penalty for additional parameters.

Popular Information Criteria:
The three most widely used criteria are:

- Akaike Information Criterion (AIC):

$$AIC = \ln(\hat{\sigma}^2) + \frac{2k}{T}, \tag{46}$$

- Schwarz's Bayesian Information Criterion (SBIC):

$$SBIC = \ln(\hat{\sigma}^2) + \frac{k}{T} \ln T, \tag{47}$$

- Hannan-Quinn Information Criterion (HQIC):

$$HQIC = \ln(\hat{\sigma}^2) + \frac{2k}{T} \ln(\ln T), \tag{48}$$

where $k = p + q + 1$ represents the number of parameters, and T is the sample size. The objective is to select the model that minimizes IC while ensuring $p \leq \bar{p}$ and $q \leq \bar{q}$.

Comparison and Practical Considerations:
Different ICs impose different penalties for model complexity:

- SBIC applies a stronger penalty term than AIC.
- If different ICs suggest different model orders:
  - SBIC is strongly consistent but may be inefficient.
  - AIC is not consistent and tends to favor larger models.

In practical applications, SBIC is often preferred for its consistency, while AIC is used when predictive accuracy is the primary goal.

4.9 ARIMA Models

The Autoregressive Integrated Moving Average (ARIMA) model extends the ARMA model by incorporating differencing to achieve stationarity. The "I" in ARIMA stands for **Integrated**, referring to the differencing operation applied to make a non-stationary time series stationary.

4.9.1 Definition

An ARIMA(p, d, q) model consists of:

- p: the order of the autoregressive (AR) component.
- d: the number of times the series is differenced to achieve stationarity.
- q: the order of the moving average (MA) component.

Mathematically, an ARIMA(p, d, q) process is expressed as:

$$\phi(L)(1 - L)^d y_t = \theta(L) u_t, \tag{49}$$

where:

- L is the lag operator, $Ly_t = y_{t-1}$.
- $(1-L)^d y_t$ represents the differencing operation.
- $\phi(L)$ and $\theta(L)$ are the AR and MA polynomials, respectively.
- u_t is a white noise error term.

4.9.2 Interpretation and Use

- A non-stationary time series may have a characteristic root on the unit circle.
- To remove this non-stationarity, the series is differenced d times until it becomes stationary.
- Once stationary, an ARMA(p, q) model can be applied to the differenced series.
- Thus, an ARMA(p, q) model applied to a differenced series is equivalent to an ARIMA(p, d, q) model applied to the original data.

4.10 Univariate Time Series Models in Practice

Univariate time series models focus on modeling a single time-dependent variable based on its past values. The primary objective is to estimate the conditional expectation $E[y_t|F_{t-1}]$, where F_{t-1} represents the historical information available up to time $t-1$.

4.10.1 The ARIMA Class of Models

The ARIMA (AutoRegressive Integrated Moving Average) class of models provides a flexible framework for modeling univariate time series data. The key types within this class include:

- **AutoRegressive (AR)** models: Use past values of the time series to predict future values.
- **Moving Average (MA)** models: Incorporate past forecast errors to improve predictions.
- **AutoRegressive Moving Average (ARMA)** models: Combine both AR and MA components.
- **AutoRegressive Integrated Moving Average (ARIMA)** models: Extend ARMA models by introducing differencing to handle non-stationary data.

4.10.2 Application in Financial Time Series

The ARIMA framework is widely used for modeling financial variables, leveraging only past values of the series and potentially past values of an error term. The key benefits of ARIMA models include:

- Capturing empirically relevant features of observed data.
- Handling non-stationary time series through differencing.

4.10.3 Use Cases

ARIMA models are particularly useful when:

- Explanatory variables are not observable or measurable.
- The explanatory variables operate at different frequencies from the target variable.
- Out-of-sample forecasting is required for future predictions.

Univariate time series modeling, particularly through ARIMA, is essential in financial forecasting, economic analysis, and various applications requiring time-dependent predictions.

4.11 Forecasting Procedures

Forecasting is a crucial aspect of time series analysis, where future values are predicted based on past observations. Effective forecasting involves evaluating both in-sample and out-of-sample performance and selecting the appropriate method for multi-period forecasting.

4.11.1 In-Sample vs. Out-of-Sample Forecasting

In time series forecasting, performance assessment can be categorized as:

- **In-sample forecasting:** Evaluates model performance within the training dataset.
- **Out-of-sample forecasting:** Tests the model on unseen data, providing a better estimate of real-world performance.

A common method for assessing forecasting accuracy is **time series cross-validation** with an expanding window, where:

- Training data grows over time.
- Testing data is sequentially evaluated at each step.

4.11.2 Multi-Period Forecasting Approaches

For forecasting multiple periods ahead, three main strategies are commonly used:

- **Scaled approach:** Adjusts forecasts based on estimated scale factors to account for long-term trends.
- **Iterated approach:** Uses a one-step-ahead forecast iteratively, where each predicted value is fed into the model to predict the next step.
- **Direct approach:** Builds separate models for each forecast horizon, reducing error accumulation.

Each of these approaches has advantages and is selected based on the nature of the dataset and forecasting objectives.

4.11.3 Forecasting ARMA models:

h-Steps Ahead Forecast ($\hat{y}_{T+h|T}$):
The model for y_{T+h} is:

$$y_{T+h} = c + \sum_{i=1}^{p} \phi_i y_{T+h-i} + \sum_{j=1}^{q} \theta_j u_{T+h-j} + u_{T+h}$$

Taking the conditional expectation:

$$\begin{aligned}\hat{y}_{T+h|T} &= E[y_{T+h}|I_T] \\ &= c + \sum_{i=1}^{p} \phi_i E[y_{T+h-i}|I_T] + \sum_{j=1}^{q} \theta_j E[u_{T+h-j}|I_T] + E[u_{T+h}|I_T]\end{aligned}$$

Applying the rules:

- $E[y_{T+h-i}|I_T] = \begin{cases} y_{T+h-i} & \text{if } T+h-i \leq T \\ \hat{y}_{T+h-i|T} & \text{if } T+h-i > T \end{cases}$
- $E[u_{T+h-j}|I_T] = \begin{cases} u_{T+h-j} & \text{if } T+h-j \leq T \\ 0 & \text{if } T+h-j > T \end{cases}$
- $E[u_{T+h}|I_T] = 0$.

So, the h-steps ahead forecast is:

$$\hat{y}_{T+h|T} = c + \sum_{i=1}^{p} \phi_i E[y_{T+h-i}|I_T] + \sum_{j=1}^{q} \theta_j E[u_{T+h-j}|I_T] \tag{50}$$

Specifically, for $h > q$: all terms $E[u_{T+h-j}|I_T]$ become 0 because $T + h - j > T$ for all $j \in [1, q]$. In this case ($h > q$), the forecast becomes purely autoregressive:

$$\hat{y}_{T+h|T} = c + \sum_{i=1}^{p} \phi_i E[y_{T+h-i}|I_T] \quad (\text{for } h > q) \tag{51}$$

The $E[y_{T+h-i}|I_T]$ terms are replaced by their actual values if $T + h - i \leq T$, or by their previously computed forecasts $\hat{y}_{T+h-i|T}$ if $T + h - i > T$.

4.12 Forecast Evaluation

Evaluating the accuracy of forecasts is essential for assessing the performance of time series models. Various error metrics provide different insights into forecast quality, considering factors like sensitivity to outliers and interpretability.

4.12.1 Error Metrics

The most common forecast evaluation metrics include:

Mean Squared Error (MSE)

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{52}$$

Pros:

- Penalizes large errors more than small ones.
- Differentiable, making it useful for optimization.

Cons:

- Sensitive to outliers.
- Not easily interpretable due to squared units.

Root Mean Squared Error (RMSE)

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \tag{53}$$

Pros:

- Same unit as the target variable, improving interpretability.

Cons:

- Still sensitive to large errors.

Mean Absolute Error (MAE)

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{54}$$

Pros:

- Less sensitive to outliers than MSE.
- Easy to interpret.

Cons:

- Not differentiable at zero, which can pose optimization issues.

Mean Absolute Percentage Error (MAPE)

$$MAPE = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \tag{55}$$

Pros:

- Provides a relative error measure, making it useful for comparing forecasts across different scales.

Cons:

- Undefined when $y_i = 0$.
- Biased towards underestimation.

Theil's U-Statistic

$$\text{Theil's U} = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{\sum y_i^2}} \tag{56}$$

Pros:

- Measures relative forecasting performance.
- Values closer to zero indicate better accuracy.

Cons:

- Less commonly used than other metrics.

4.12.2 Comparison of Error Metrics

Different metrics are suitable for different forecasting contexts. RMSE is widely used for its interpretability, while MAPE is useful when comparing forecast errors across different scales. Theil's U-statistic provides a relative measure of forecasting performance.

Metric	Scale	Outlier Sensitivity	Interpretability
MSE	Squared	High	Low
RMSE	Same as target	High	Moderate
MAE	Same as target	Low	High
MAPE	Percentage	Medium	High
Theil's U	Relative	Medium	Moderate

Table 1: Comparison of Forecast Evaluation Metrics

4.13 Forecasting Comparisons

Forecasting comparisons are essential in evaluating the accuracy and reliability of different forecasting models. This section presents key statistical tests used to compare forecasts, along with considerations for economic significance.

4.13.1 Diebold-Mariano Test

The Diebold-Mariano (DM) test is used to compare the accuracy of two competing forecast models. It is defined as:

$$DM = \frac{\bar{d}}{\sqrt{\text{Var}(d)/n}}, \tag{57}$$

where $\bar{d}$ represents the mean loss differential, and $\text{Var}(d)$ denotes its variance.

Pros:

- Allows for direct comparison of forecast accuracy between models.
- Accounts for autocorrelation in forecast errors.

Cons:

- Requires sufficiently long samples for reliable inference.
- Sensitive to the choice of loss function.

4.13.2 Clark-West Test

The Clark-West (CW) test is specifically designed for comparing nested models. It adjusts for the bias that arises when comparing forecasts from nested models. The CW test statistic is given by:

$$CW = \frac{\bar{d} - \lambda}{\sqrt{\text{Var}(d)/n}}, \tag{58}$$

where λ is an adjustment term accounting for nested model comparisons.

Pros:

- Designed for comparing nested models.

- Corrects for bias in nested model comparisons.

Cons:

- More complex than the Diebold-Mariano test.
- Sensitive to assumptions about error variance.

4.13.3 Statistical vs. Economic Loss Functions

While statistical accuracy is important in forecasting comparisons, it does not necessarily translate into economic significance. It is crucial to document both dimensions of empirical findings:

- Statistical accuracy $\neq$ economic significance.
- Consideration of economic impact is necessary for practical decision-making.

5 Guest Lecture Hafslund (Not in Book)

5.1 The Power Market Context

5.1.1 Market Definition and Price Determination

A **power market** is a marketplace where electricity is bought and sold between producers and consumers. A fundamental characteristic of electricity is that supply and demand must be balanced continuously to maintain grid stability.

In many European markets, prices are determined in a **Day-Ahead Market**. The price formation typically follows the **merit order principle**: available generation sources are stacked from lowest to highest marginal cost. The market clearing price is set by the marginal cost of the last unit of generation needed to meet demand in a given hour. This process is repeated for each hour, leading to dynamic and often volatile prices.

5.1.2 Heterogeneity of Assets and Revenue Streams

Even though participants operate within the same market, different generation assets (e.g., hydropower, wind, solar, thermal) obtain different revenues. This is due to:

- **Production Profiles:** Renewable sources like wind and solar have intermittent and variable production patterns dictated by weather conditions. Hydropower production depends on water inflows and reservoir management.
- **Flexibility:** Assets differ in their ability to adjust output quickly in response to price signals or system needs. Hydropower with reservoirs, for instance, offers significant flexibility.

This heterogeneity underscores the need for sophisticated forecasting not just for market prices, but also for production levels of various asset types and key input factors like hydrological inflows.

5.2 The Role of Forecasting and Modelling

Forecasting and quantitative modelling are indispensable tools in power markets, supporting a wide array of decisions across different time horizons.

5.2.1 Key Decision Areas

1. **Production Planning (Short-to-Medium Term):** Forecasts of prices, inflows (for hydropower), and renewable generation are used to optimize the dispatch of generation assets. For example, determining the optimal use of water from reservoirs to maximize revenue while ensuring supply security.
2. **Risk and Hedging Analysis / Trading (Short-to-Long Term):** Models help assess the future value of generation portfolios and inform hedging strategies to secure revenues at certain price levels. This involves forecasting price distributions and correlations.
3. **Investment Analysis (Long Term):** Decisions on whether to invest in new generation capacity (e.g., offshore wind, solar, new hydropower plants, flexibility enhancements) are heavily reliant on long-term price forecasts and scenarios for power system development. These models build business cases by projecting future revenues and costs.

5.2.2 Objective of Market Models

Many market models aim to simulate the behavior of electricity markets. They often incorporate an **objective function**, such as:

- Maximization of social welfare.
- Minimization of total system operational costs.

These models solve for equilibrium prices and dispatch decisions given a set of inputs.

5.3 Key Variables for Forecasting

Effective participation and planning in power markets require forecasting several critical variables:

- **Electricity Prices:** Spot prices (e.g., Day-Ahead) and forward prices.
- **Renewable Generation:** Production from wind power (onshore/offshore) and solar power, which are weather-dependent.

- **Hydropower Inflows:** The amount of water flowing into reservoirs, crucial for hydro-dominant systems.
- **Demand:** Electricity consumption patterns.
- **Fuel Prices:** Prices of gas, coal, CO2 emission allowances, which influence the marginal costs of thermal power plants.

5.4 Types of Models Employed

A variety of modelling approaches are used, often categorized as fundamental or statistical.

5.4.1 Fundamental Models

These models are based on the economic and physical principles governing the power market. They simulate the market clearing process by optimizing generation dispatch to meet demand, considering network constraints and plant characteristics.

- **Deterministic Models:** Assume all input variables (demand, fuel prices, renewable availability, etc.) are known with certainty for the modelling period. They perform dispatch optimization under these known conditions.
- **Stochastic Models:** Explicitly incorporate uncertainty in key input variables (e.g., renewable generation, inflows, outages). They often use probabilistic methods or scenario analysis to determine robust dispatch or investment strategies.

Fundamental models are crucial for understanding market dynamics, policy impacts, and long-term investment analysis, especially when considering structural market changes (e.g., energy transition).

5.4.2 Statistical Models (Econometric / Machine Learning)

These models leverage historical data to identify patterns and relationships, which are then used to forecast future values.

- **Regression-based models:** Including time series models like ARIMAX (Autoregressive Integrated Moving Average with eXogenous variables) and more complex variants like LEAR (LASSO-Estimated AutoRegressive model, a parameter-rich ARX model estimated using LASSO for variable selection).
- **Machine Learning (ML) models:** Such as Artificial Neural Networks (ANNs), support vector machines, gradient boosting, etc. These are increasingly used, particularly for short-term forecasting of renewable generation, inflows, and prices, given their ability to capture complex non-linear relationships.

Statistical models are often favored for short-term operational forecasting due to their adaptability and performance on high-frequency data.

5.5 Practical Considerations in Model Building and Selection

The choice and implementation of a model involve several practical trade-offs and considerations.

5.5.1 Method Selection

- **Complexity vs. Performance:** More complex models (e.g., deep neural networks) do not always outperform simpler ones (e.g., linear regression, naive models), especially with limited data or for specific tasks. The trade-off between model complexity (e.g., number of inputs, model depth/parameters) and prediction error (e.g., RMSE) is a key concern.
- **Short-term vs. Long-term Needs:**
  - For **short-term trends** (e.g., next-day price forecasting), simpler statistical models leveraging recent data might be sufficient.
  - For capturing **structural market changes** (e.g., impact of increased renewable penetration, policy shifts for the green transition), fundamental models or more sophisticated hybrid approaches are often more adequate.

5.5.2 Implementation Challenges

- **Forecast Frequency and Refitting:** Some complex models may have long training times, making them unsuitable if continuous forecasts and frequent model refitting with new information are required.
- **Computational Power:** Models with millions of parameters or those requiring extensive simulations (e.g., stochastic fundamental models) demand significant computational resources. Efficiency is a key factor.
- **Accuracy vs. Computation Time:** A balance must be struck. While higher accuracy is desirable, the computational cost to achieve marginal gains might be prohibitive for certain applications.
- **Data Availability and Quality:** ML and neural network models, in particular, may require large, high-quality datasets to outperform simpler models and generalize well.

5.6 Modelling Uncertainty

Uncertainty is an inherent feature of power markets and affects all decisions, regardless of the time horizon. Key sources of uncertainty include weather-driven renewable generation, hydrological conditions, fuel price volatility, and demand fluctuations.

- It is crucial to **model this uncertainty** rather than relying solely on deterministic point forecasts. This can involve:
  - Probabilistic forecasting (generating prediction intervals or full predictive distributions).
  - Scenario analysis.
  - Stochastic optimization in fundamental models.
- For example, when managing hydropower reservoirs, it's vital to consider the range of possible future inflows ("possibilities") rather than just a single expected value, to make robust operational decisions.

5.7 Utility of Models Beyond Operational Decisions

While models are essential for day-to-day operations and trading, their utility extends to strategic and policy-related questions.

- They help analyze "what if" scenarios, assessing the potential impacts of new regulations, market designs, or technological advancements.
- The famous adage, *"All models are wrong, but some are useful,"* underscores their value: they provide structured frameworks for understanding complex systems and making informed decisions, even if they are imperfect representations of reality.

6 Volatility Models

6.1 Definition: Volatility

Volatility measures the dispersion of returns for a given asset or portfolio and reflects the level of uncertainty or risk associated with price movements. It is commonly defined as the standard deviation of returns, making it the second moment of the conditional return distribution. Since volatility is not directly observable, it must be estimated using historical data, leading to estimation errors.

6.2 Stylized Facts of Financial Returns

Financial returns exhibit several empirical properties that have been widely documented in the literature. These are known as the stylized facts of financial returns:

Key Stylized Facts:

- **Nonstationarity of Prices, Stationarity of Returns:** Asset prices P_t follow a random walk, making them nonstationary. However, returns r_t are generally stationary.
- **Absence of Autocorrelation in Returns:** Returns ϵ_t exhibit little to no serial correlation, behaving like white noise.
- **Autocorrelation in Squared Returns:** While raw returns are uncorrelated, squared returns ϵ_t^2 or absolute returns $|\epsilon_t|$ show significant positive autocorrelation, often modeled as weak white noise.
- **Volatility Clustering:** Periods of high volatility tend to be followed by high volatility, and periods of low volatility tend to be followed by low volatility. This means that large absolute returns are likely to be followed by other large absolute returns.
- **Fat-Tailed Distribution:** The empirical distribution of returns exhibits excess kurtosis (leptokurtic), meaning it has fatter tails than a normal distribution, indicating a higher likelihood of extreme returns.
- **Leverage Effect:** Negative returns (price decreases) tend to increase volatility more than positive returns (price increases). This asymmetric volatility response is a key feature of financial time series.

6.3 Empirical vs. Implied Volatility

There are two primary ways to measure volatility:

- **Empirical volatility** (under the physical measure $\mathbb{P}$), derived from historical data.
- **Implied volatility** (under the risk-neutral measure $\mathbb{Q}$), inferred from market prices of options.

6.3.1 Empirical Volatility ($\mathbb{P}$-measure)

Empirical (or historical) volatility is computed using past asset prices and follows from time-series models in empirical finance.

Definition: Consider a covariance stationary time series $\{y_t\}_{t=1}^T$, where:

$$E[y_t] = \mu, \quad \text{Var}[y_t] = \sigma^2 \quad \text{(constant over time)}.$$

The conditional mean of the process at time t is:

$$\mu_t = E[y_t|\mathcal{I}_{t-1}],$$

where $\mathcal{I}_t$ represents the information set available at time t.

The conditional variance (volatility) is then defined as:

$$\sigma_t^2 = \text{Var}[y_t|\mathcal{I}_{t-1}].$$

Thus, empirical volatility depends on the available past information set $\mathcal{I}_t$.

6.3.2 Implied Volatility ($\mathbb{Q}$-measure)

Implied volatility is a forward-looking measure, derived from option prices under the risk-neutral measure $\mathbb{Q}$. It represents the market's expectations of future volatility.

6.3.3 Differences Between Empirical and Implied Volatility

Empirical and implied volatility often differ due to risk premia and market expectations. The variance risk premium (VRP) is the difference between the expectations of variance under the risk-neutral and physical measures:

$$VRP_{t,T} = E_t^{\mathbb{Q}}(V_{t,T}^2) - E_t^{\mathbb{P}}(V_{t,T}^2),$$

where:

- $V_{t,T}^2$ refers to return variation.
- $E_t^{\mathbb{Q}}(V_{t,T}^2)$ is the ex-ante forecast of variance under the risk-neutral measure, derived from implied volatility.
- $E_t^{\mathbb{P}}(V_{t,T}^2)$ is the ex-ante forecast of variance under the physical measure, derived from historical data.

6.4 Models for Empirical Volatility

6.4.1 Estimating Empirical Volatility

Empirical volatility, denoted as $E_t^{\mathbb{P}}(V_{t,T}^2)$, can be estimated using various statistical models. The estimation methods are generally divided into two main approaches:

- **Time series models:** The volatility process σ_t^2 is treated as a *deterministic* function of past observations.
- **Stochastic volatility models:** The volatility process σ_t^2 follows its own *stochastic* evolution, capturing more complex market dynamics.

This course will focus on time series models for estimating volatility.

6.5 GARCH Models

6.5.1 Engle's ARCH Model

In order to capture the empirical phenomenon of *volatility clustering*, Robert Engle (1982) introduced the *Autoregressive Conditional Heteroskedasticity (ARCH)* model. This model laid the foundation for nonlinear modeling of financial time series, particularly in estimating time-varying volatility.

The ARCH(1) Model:

The ARCH(1) model assumes that the return process ϵ_t follows:

$$\epsilon_t = \sqrt{\sigma_t^2} z_t, \quad z_t \sim D(0,1),$$

where D is a distribution with mean 0 and variance 1. The conditional variance σ_t^2 is modeled as:

$$\sigma_t^2 = \omega + \alpha \epsilon_{t-1}^2, \quad \omega > 0, \quad \alpha \geq 0.$$

Key Properties:

- The coefficient α must satisfy certain conditions to ensure the process remains well-behaved, meaning it has finite variance and is stationary.
- Intuitively, the conditional variance of returns depends on the past squared return. If past returns were large (either positive or negative), future volatility is expected to be higher.

The ARCH(q) Model:

The ARCH(1) model can be generalized to an *ARCH(q) model*, which allows for a longer memory in past squared returns:

$$\sigma_t^2 = \omega + \alpha_1 \epsilon_{t-1}^2 + \alpha_2 \epsilon_{t-2}^2 + \cdots + \alpha_q \epsilon_{t-q}^2.$$

where:

- $\omega > 0$ ensures that the variance remains positive.
- $\alpha_i \geq 0$ for all i ensures that the conditional variance is non-negative.
- The sum of α_i determines the persistence of volatility shocks in the model.

6.5.2 Detecting ARCH Effects: The ARCH-LM Test

To determine whether a time series exhibits *ARCH effects*, Engle (1982) proposed the ARCH-LM (Lagrange Multiplier) test. This test is widely used to detect conditional heteroskedasticity in financial time series.

The ARCH-LM Test Procedure:

The test follows these steps:

1. **Estimate the following autoregressive model using OLS:**

$$\epsilon_t^2 = \alpha_0 + \alpha_1 \epsilon_{t-1}^2 + \alpha_2 \epsilon_{t-2}^2 + \cdots + \alpha_q \epsilon_{t-q}^2 + u_t.$$

2. **Formulate the null hypothesis (H_0):**

$$H_0 : \alpha_1 = 0, \quad \alpha_2 = 0, \quad \ldots, \quad \alpha_q = 0.$$

   This implies that past squared residuals do not explain future conditional variance, meaning no ARCH effects are present.

3. **Compute the test statistic:**

$$T \cdot R^2.$$

   Here, T is the sample size, and R^2 is the coefficient of determination from the regression.

Interpretation of the Test:

Under the null hypothesis of no ARCH effects, the test statistic follows a chi-square distribution with q degrees of freedom:

$$T \cdot R^2 \sim \chi_q^2.$$

Decision Rule:

- If the computed test statistic is greater than the critical value from the χ_q^2 distribution, we reject H_0 and conclude that ARCH effects are present.
- If the test statistic is smaller than the critical value, we fail to reject H_0, implying no significant ARCH effects.

6.5.3 Limitations of ARCH

- In practice, only rather rich ARCH parameterizations are able to fit financial series adequately.
- However, largely parameterized models can be unstable in forecasting and hard to estimate.
- In order to overcome the shortcomings of the ARCH, Tim Bollerslev proposed a generalisation of the ARCH model called GARCH.
- The model allows fitting financial returns adequately while keeping the number of parameters small.
- In practice, the GARCH model is one of the most successfully employed volatility models.

6.5.4 GARCH

The GARCH(1,1) Model:
The GARCH(1,1) model is defined as:

$$\epsilon_t = \sqrt{\sigma_t^2} z_t, \quad z_t \sim D(0,1) \tag{59}$$

where D is a distribution with mean 0 and variance 1. The conditional variance equation follows:

$$\sigma_t^2 = \omega + \alpha \epsilon_{t-1}^2 + \beta \sigma_{t-1}^2 \tag{60}$$

where $\omega > 0$, $\alpha \geq 0$, and $\beta > 0$ with the stationarity condition:

$$\alpha + \beta < 1. \tag{61}$$

The GARCH(p,q) Model:
The general GARCH(p,q) model extends GARCH(1,1) as follows:

$$\epsilon_t = \sqrt{\sigma_t^2} z_t, \quad z_t \sim D(0,1) \tag{62}$$

where D is a distribution with mean 0 and variance 1, and the conditional variance equation is:

$$\sigma_t^2 = \omega + \sum_{i=1}^{q} \alpha_i \epsilon_{t-i}^2 + \sum_{j=1}^{p} \beta_j \sigma_{t-j}^2. \tag{63}$$

The constraints on the parameters are:

$$\omega > 0, \quad \alpha_i \geq 0, \quad \beta_j > 0, \quad \sum \alpha_i + \sum \beta_j < 1. \tag{64}$$

Stationarity Condition:
A necessary condition for weak stationarity of the GARCH(1,1) model is that the process has a finite, time-invariant unconditional variance. That is,

$$\alpha + \beta < 1. \tag{65}$$

To derive this, we first take the expectation on both sides of the variance equation:

$$\mathbb{E}[\sigma_t^2] = \mathbb{E}[\omega + \alpha \epsilon_{t-1}^2 + \beta \sigma_{t-1}^2]. \tag{66}$$

Since ω is a constant,

$$\mathbb{E}[\sigma_t^2] = \omega + \alpha \mathbb{E}[\epsilon_{t-1}^2] + \beta \mathbb{E}[\sigma_{t-1}^2]. \tag{67}$$

We want to find an expression for $\mathbb{E}[\epsilon_t^2]$
Squaring ϵ_t, gives us

$$\epsilon_t^2 = \sigma_t^2 z_t^2. \tag{68}$$

Taking the unconditional expectation:

$$\mathbb{E}[\epsilon_t^2] = \mathbb{E}[\sigma_t^2 z_t^2]. \tag{69}$$

$$\mathbb{E}[\epsilon_t^2] = \mathbb{E}[\sigma_t^2]. \tag{70}$$

Thus, defining $\mathbb{E}[\sigma_t^2] = \sigma^2$, we get:

$$\sigma^2 = \omega + \alpha\sigma^2 + \beta\sigma^2. \tag{71}$$

Rearranging,

$$\sigma^2(1 - \alpha - \beta) = \omega. \tag{72}$$

For the variance to be finite and positive, we require $\omega > 0$ and $\alpha + \beta < 1$. This ensures stationarity, meaning that the process does not exhibit an explosive behavior where the variance grows indefinitely.

Unconditional Variance:
The unconditional variance is the long-run average variance of the process, as opposed to the conditional variance, which depends on past information. If $\alpha + \beta < 1$, then the unconditional variance of GARCH(1,1) exists and is given by:

$$\mathbb{E}[\epsilon_t^2] = \sigma^2 = \frac{\omega}{1 - \alpha - \beta}, \quad \text{if } \omega > 0. \tag{73}$$

This result follows directly from solving the stationarity equation derived earlier. If $\alpha + \beta \geq 1$, the denominator approaches zero or becomes negative, leading to an infinite or undefined variance, implying a non-stationary process.

For a general GARCH(p,q) model, the unconditional variance follows a similar derivation:

$$\mathbb{E}[\epsilon_t^2] = \sigma^2 = \frac{\omega}{1 - \sum_{i=1}^{q} \alpha_i - \sum_{j=1}^{p} \beta_j}. \tag{74}$$

Again, the stationarity condition ensures that the denominator remains positive, which is necessary for the variance to be finite.

6.5.5 Limitations of GARCH

The simple GARCH(p, q) has some limitations. The most important is that it cannot take into account the dependence between volatility and the sign of past returns.

- Standard GARCH models assume that positive and negative error terms have a symmetric effect on the volatility, i.e., good and bad news have the same effect on the volatility.
- In many real situations, the volatility reacts asymmetrically to the sign of the shocks. In particular, negative past returns have a bigger effect on σ_t^2 than positive returns of the same size.

- This dependence is due to the leverage effect, i.e., a negative shock to returns would increase the debt-to-equity ratio, which in turn will increase the uncertainty of future returns.

The Leverage Effect and the News Impact Curve:

Economic Explanation of the Leverage Effect:
The leverage effect describes the observed negative correlation between stock returns and changes in volatility. It is seen that this effect is linked to a firm's financial leverage: when stock prices decline, the firm's equity capital is reduced, increasing its leverage and consequently its risk level.

A higher financial leverage implies greater sensitivity of equity returns to firm performance, leading to increased volatility. This effect is particularly relevant to investors, as it suggests that volatility is not constant but dynamically responds to past returns. However, it may be argued that financial leverage alone does not fully explain the leverage effect, suggesting that other factors also contribute to this phenomenon.

The News Impact Curve (NIC):
The News Impact Curve (NIC) quantifies the empirical relationship between news shocks and return volatility. It provides a framework for understanding how new information impacts market uncertainty.

The NIC illustrates how past return shocks influence the conditional volatility in financial markets, an essential feature in volatility modeling. In a Generalized Autoregressive Conditional Heteroskedasticity (GARCH) context, the NIC typically exhibits an asymmetric U-shaped pattern, reflecting the leverage effect: negative shocks tend to increase volatility more than positive shocks of the same magnitude.

6.5.6 The GJR (1,1) Model (TGARCH model)

The GJR model, introduced by Glosten, Jagannathan, and Runkle (1993), extends the GARCH model to account for leverage effects, where negative shocks tend to increase volatility more than positive shocks of the same magnitude. The general form of a GJR(p,q) model is given by the following equation for the conditional variance:

$$\sigma_t^2 = \omega + \sum_{i=1}^{q} \alpha_i u_{t-i}^2 + \sum_{j=1}^{p} \beta_j \sigma_{t-j}^2 + \sum_{i=1}^{q} \gamma_i u_{t-i}^2 I(u_{t-i} < 0) \tag{75}$$

where $I(A)$ is an indicator function that takes the value 1 if A is verified and 0 otherwise. The indicator function $I(u_{t-i} < 0)$ captures asymmetry in the conditional variance, as it ensures that negative past shocks contribute differently to volatility compared to positive ones.

Special Case: GARCH Model:

If $\gamma_i = 0$ for all i, the model reduces to the standard GARCH model, which does not account for leverage effects.

The GJR (1,1) Model:
A specific case of the GJR(p,q) model with $p = 1, q = 1$ is given by:

$$\sigma_t^2 = \omega + \alpha u_{t-1}^2 + \beta \sigma_{t-1}^2 + \gamma u_{t-1}^2 I(r_{t-1} < 0) \tag{76}$$

where r_{t-1} represents past returns.

Properties of the GJR (1,1) Model:

- **Positivity Conditions:** To ensure the conditional variance remains non-negative, the following conditions must hold:

$$\omega > 0, \quad \alpha \geq 0, \quad \beta \geq 0, \quad (\alpha + \gamma) \geq 0. \tag{77}$$

- **Covariance Stationarity:** Under the assumption that z_t follows a symmetric distribution, the process is covariance stationary if and only if:

$$\alpha + \beta + \frac{\gamma}{2} < 1. \tag{78}$$

- **Relation to GARCH(1,1):** The GJR(1,1) model can be rewritten as a GARCH(1,1) model with time-varying parameters, making it a more flexible tool for modeling financial volatility dynamics.

This model is particularly useful in capturing the asymmetric effects of market shocks, making it relevant for risk management and financial modeling.

6.5.7 The Exponential GARCH (EGARCH) model

The Exponential GARCH (EGARCH) model, introduced by Nelson (1991), extends the standard GARCH framework by allowing the volatility to depend on both the magnitude and sign of lagged residuals. This model captures asymmetric volatility effects, where negative shocks may impact volatility differently than positive shocks.

The general form of the EGARCH model of order (p, q) is given by:

$$\log\left(\sigma_t^2\right) = \omega + \sum_{i=1}^{q} \left[\alpha_i \left| \frac{u_{t-i}}{\sigma_{t-i}} \right| + \gamma_i \frac{u_{t-i}}{\sigma_{t-i}} \right] + \sum_{j=1}^{p} \beta_j \log\left(\sigma_{t-j}^2\right) \tag{79}$$

where:

- ω is a constant,
- α_i represents the magnitude effect,

- γ_i captures the asymmetric effect; if $\gamma_i < 0$, negative shocks increase volatility more than positive shocks,
- β_j accounts for persistence in volatility.

This model allows volatility to react asymmetrically to good and bad news, making it particularly useful in financial modeling where leverage effects are present.

A key advantage of the EGARCH model is that the logarithmic specification of variance guarantees positivity of σ_t^2 without requiring parameter constraints. This makes it more flexible in incorporating explanatory variables in the volatility equation.

For an EGARCH(1,1) model, the specification simplifies to:

$$\log\left(\sigma_t^2\right) = \omega + \alpha \left|\frac{u_{t-1}}{h_{t-1}}\right| + \gamma \frac{u_{t-1}}{h_{t-1}} + \beta \log\left(\sigma_{t-1}^2\right) \tag{80}$$

where:

- The term $\log(\sigma_t^2)$ ensures non-negativity of the conditional variance.
- The asymmetric effect is captured by γ, where $\gamma < 0$ implies leverage effects, meaning that negative shocks lead to higher volatility than positive shocks of the same magnitude.

6.5.8 Impact of good and bad news: GJR and EGARCH

- The GJR allow only bad news to have a different impact on volatility;
- The EGARCH allow good and bad news to have a different impact on volatility;
- The EGARCH allow big news to have a greater impact on volatility than the GARCH and GJR model. This is because the exponential dominates the quadratic relation.

6.5.9 Asymmetric Power ARCH (APARCH)

The Asymmetric Power ARCH (APARCH) model extends traditional ARCH/GARCH models by allowing for power transformations of conditional variance and incorporating asymmetry in the response to shocks.

Conditional Variance Equation:
The conditional variance of an APARCH model is given by:

$$\sigma_t^\delta = \omega + \sum_{i=1}^{q} \alpha_i \left(|u_{t-i}| - \gamma_i u_{t-i}\right)^\delta + \sum_{j=1}^{p} \beta_j \sigma_{t-j}^\delta \tag{81}$$

where the parameters satisfy the conditions:

- $\omega > 0$, $\alpha_i \geq 0$, $\beta_j \geq 0$, $\delta > 0$,
- $|\gamma_i| < 1$.

Special Cases of the APARCH Model: By choosing specific values for the coefficients, the APARCH model generalizes several well-known GARCH-type models:

- The standard GARCH model if $\delta = 2$ and $\gamma_i = 0, \forall i$.
- The GJR-GARCH model if $\delta = 2$.
- The Threshold GARCH (T-GARCH) of Zakoian if $\delta = 1$. (See later)

6.5.10 The GARCH in Mean (GARCH-M) Model

The Risk-Return Relationship:
Financial theory suggests the presence of a positive relationship between the mean (excess) return and the conditional variance of returns. This relationship is particularly relevant in the framework of the Capital Asset Pricing Model (CAPM), where it is assumed to be linear.

Since the conditional variance serves as a measure of risk, assets with higher risk levels tend to yield higher expected returns. This results in the mean return being dependent on the riskiness of the asset, which is commonly referred to as the risk premium.

Mathematical Formulation of the GARCH-M Model:

The Generalized Autoregressive Conditional Heteroskedasticity in Mean (GARCH-M) model extends the standard GARCH framework by allowing the conditional mean of the return process to depend explicitly on the conditional variance. This feature is particularly useful in financial modeling, where risk (measured by variance) can influence expected returns.

A GARCH(1,1)-M model is specified as follows:

$$r_t = X_t' b + \delta \sigma_t^2 + u_t, \tag{82}$$

$$u_t = \sigma_t z_t, \tag{83}$$

$$\sigma_t^2 = \omega + \alpha_1 u_{t-1}^2 + \beta_1 \sigma_{t-1}^2. \tag{84}$$

Explanation of Variables and Parameters:

- r_t: The return at time t, which is composed of a deterministic part and a stochastic part.
- $X_t (k \times 1)$: A vector of exogenous regressors that influence the mean of the return process.
- $b (k \times 1)$: A vector of unknown parameters associated with the regressors.

- δ: The risk premium coefficient, capturing the impact of conditional variance σ_t^2 on the expected return.
- u_t: The error term representing the innovation (shock) in returns at time t, which is assumed to be conditionally heteroskedastic.
- σ_t^2: The conditional variance of r_t, which evolves dynamically over time.
- z_t: A standard normal i.i.d. process, i.e., $z_t \sim \mathcal{N}(0,1)$.
- ω: The intercept term in the variance equation, ensuring the process remains strictly positive.
- α_1: The ARCH parameter, measuring the sensitivity of conditional variance to past squared shocks u_{t-1}^2.
- β_1: The GARCH parameter, capturing the persistence of past conditional variance σ_{t-1}^2.

Interpretation of the Model Equations:

- The first equation defines the return process, where the conditional mean $\mathbb{E}[r_t|\mathcal{F}_{t-1}]$ depends on both the exogenous variables X_t and the conditional variance σ_t^2, reflecting a risk-return tradeoff.
- The second equation specifies the error structure, where u_t is scaled by the time-varying standard deviation σ_t.
- The third equation is the conditional variance equation, modeling volatility as a function of past squared innovations and past conditional variance.

Rearranging the return equation, we define the conditional mean of returns explicitly as:

$$\mu_t = X_t' b + \delta \sigma_t^2. \tag{85}$$

This formulation illustrates that expected returns are influenced by risk, as captured by the term $\delta\sigma_t^2$. If $\delta > 0$, higher risk (variance) is associated with higher expected returns, aligning with traditional asset pricing theories.

The model is useful in financial econometrics for capturing time-varying volatility and analyzing risk premia in asset returns.

Extensions and Generalizations:
Several generalizations and refinements of the GARCH-M model have been proposed:

- The model can be extended to account for non-linear dependence on σ_t^2, such as a linear dependency in log-volatility. Additionally, σ_t^2 is sometimes replaced by the conditional standard deviation instead.

- Alternative models for conditional variance can be employed, including GJR-GARCH, EGARCH, and other variants.
- Since σ_t^2 appears in the mean equation, estimating the model's parameters consistently requires simultaneous estimation of both the mean and conditional variance parameters.

6.5.11 Integrated GARCH (IGARCH)

The *Integrated GARCH* (IGARCH) model is a special case of the Generalized Autoregressive Conditional Heteroskedasticity (GARCH) framework in which the effects of volatility shocks do not decay over time but instead persist indefinitely. This permanent persistence arises when the sum of the ARCH and GARCH parameters equals one.

Model Specification Consider the standard GARCH(p, q) model:

$$\sigma_t^2 = \omega + \sum_{i=1}^{p} \alpha_i \epsilon_{t-i}^2 + \sum_{j=1}^{q} \beta_j \sigma_{t-j}^2. \tag{86}$$

If the following constraint holds:

$$\sum_{i=1}^{p} \alpha_i + \sum_{j=1}^{q} \beta_j = 1, \tag{87}$$

then the model is said to be *integrated*, or IGARCH. This implies that any shock to the conditional variance σ_t^2 will have a lasting (permanent) effect, analogous to a unit-root process in standard time series analysis.

Implications:

- Since the conditional variance follows a unit root process, shocks to volatility are highly persistent.
- The unconditional variance of the process is not defined, meaning the model does not revert to a finite long-run variance.
- IGARCH models are commonly used in financial time series when volatility clustering exhibits long memory effects.

The presence of a unit root in the variance equation suggests that volatility shocks do not decay over time, making IGARCH particularly relevant for modeling financial markets with persistent volatility dynamics.

6.5.12 FIGARCH (Fractionally Integrated GARCH)

The **Fractionally Integrated GARCH (FIGARCH)** model, extends the standard GARCH framework to account for long memory in conditional variance.

Purpose:

Standard GARCH models imply an exponential decay of the influence of past squared innovations on current conditional variance. While Integrated GARCH (IGARCH) models allow for infinite persistence ($\alpha + \beta = 1$), they still exhibit exponential decay of the *impact* of a single shock. Empirical evidence from financial markets frequently shows a much slower, hyperbolic decay of volatility shocks. FIGARCH models are designed to capture this observed *long-range dependence* or *long memory* in volatility, providing a more realistic representation for financial time series data that exhibit highly persistent volatility clustering.

Parameters:
The most significant parameter distinguishing FIGARCH from other GARCH models is the **fractional differencing parameter** d.

- Its range is typically $0 < d < 1$.
- If $d = 0$, the model reduces to a standard GARCH process, implying short memory.
- If $d = 1$, the model reduces to an Integrated GARCH (IGARCH) process, implying permanent shocks to the conditional variance.
- A value of d between 0 and 1 signifies long memory, where shocks to volatility decay at a hyperbolic rate. This means their influence persists over a very long time but eventually dies out, unlike the infinite persistence of IGARCH.

6.6 Estimating (G)ARCH

- ARCH models are typically estimated by Maximum Likelihood
- The ML estimator has no closed form expression and needs to be found numerically
- In practice, QML is used instead of ML

6.7 Residuals' Diagnostics in GARCH Models

To assess model adequacy in GARCH estimation, we analyze the standardized residuals, defined as:

$$\hat{z}_t = \frac{\epsilon_t}{\hat{\sigma}_t} = \frac{\epsilon_t}{\sqrt{\hat{\omega} + \sum_{i=1}^{q} \hat{\alpha}_i \epsilon_{t-i}^2}}. \tag{88}$$

where:

- $\epsilon_t = y_t - \hat{\mu}_t$ is the estimated residual.
- $\hat{\sigma}_t$ is the estimated conditional standard deviation.
- $\hat{\omega}$ and $\hat{\alpha}_i$ are obtained via Maximum Likelihood Estimation (MLE) under the assumed distribution D.

Checking Model Adequacy:
A well-specified GARCH model should produce standardized residuals $\hat{z}_t$ that satisfy:

1. **Distributional Properties:** They should be approximately distributed according to the assumed distribution D (e.g., Normal, Student-t, GED).
2. **Independence:** They should exhibit no significant dependence in:
   - Levels ($\mathbb{E}[z_t z_{t-k}] \approx 0$ for $k \neq 0$).
   - Absolute values ($|\hat{z}_t|$ should be uncorrelated).
   - Squared values ($\hat{z}_t^2$ should be uncorrelated).

If the standardized residuals are not i.i.d., the model may be misspecified, suggesting the need for:

- A different error distribution.
- Higher-order GARCH terms.
- Additional exogenous variables.

Practical Residual Diagnostics:
Standard methods for checking residual behavior include:

- Histogram and QQ-plots to compare empirical distribution with D.
- Ljung-Box Test for autocorrelation in residuals and squared residuals.
- ARCH-LM Test to detect remaining volatility clustering.

A properly specified model should have residuals that resemble white noise, ensuring that all relevant dynamics have been captured.

6.8 The Heterogeneous Autoregressive (HAR) Model

The *Heterogeneous Autoregressive (HAR) model* is a regression-based approach to modeling *Realized Volatility (RV)*, designed to capture the persistent nature of volatility at different time horizons. It extends standard autoregressive models by incorporating volatility components at daily, weekly, and monthly levels.

6.8.1 HAR Model Specification

The HAR model is given by:

$$RV_t = \beta_0 + \beta_D RV_{t-1} + \beta_W RV_{t-1,t-5} + \beta_M RV_{t-1,t-22} + \epsilon_t. \tag{89}$$

where:

- RV_t is the realized volatility at time t.
- RV_{t-1} represents the daily realized volatility component.
- $RV_{t-1,t-5}$ represents the weekly (5-day) average realized volatility.
- $RV_{t-1,t-22}$ represents the monthly (22-day) average realized volatility.
- $\beta_0, \beta_D, \beta_W, \beta_M$ are regression coefficients.
- ϵ_t is an error term.

6.8.2 Estimation of the HAR Model

The HAR model is typically estimated using Ordinary Least Squares (OLS) regression, which minimizes the sum of squared errors:

$$\min_{\beta} \sum_{t=1}^{T} \left(RV_t - \beta_0 - \beta_D RV_{t-1} - \beta_W RV_{t-1,t-5} - \beta_M RV_{t-1,t-22}\right)^2. \tag{90}$$

To account for potential heteroskedasticity in the error term ϵ_t, inference is performed using heteroskedasticity and autocorrelation-consistent (HAC) standard errors, such as those introduced by Newey-West.

6.8.3 Interpretation of HAR Components

The HAR model assumes that different market participants react to volatility at different time horizons:

- **Daily volatility** (RV_{t-1}) captures short-term trader behavior.
- **Weekly volatility** ($RV_{t-1,t-5}$) reflects medium-term effects, such as fund rebalancing.
- **Monthly volatility** ($RV_{t-1,t-22}$) accounts for long-term investor decisions.

6.8.4 Advantages of the HAR Model

- **Simple and intuitive**: The HAR model extends AR models by incorporating volatility components over different time horizons.
- **Captures volatility persistence**: By including multi-scale volatility components, the HAR model better reflects how volatility propagates over time. **Empirical effectiveness**: It has been shown to outperform standard GARCH models in forecasting realized volatility.

In summary, the HAR model provides a practical and effective framework for modeling realized volatility, leveraging high-frequency data and addressing heterogeneous market reactions to volatility fluctuations.

6.9 MIDAS Models

The *Mixed Data Sampling (MIDAS) regression* is an econometric technique that allows for data sampled at different frequencies to be used within the same regression framework. Unlike traditional time series models, which require that regressors and the dependent variable share the same frequency, MIDAS enables the incorporation of *high-frequency data into low-frequency regressions* in a parsimonious yet flexible manner.

6.9.1 Motivation for MIDAS Models

Standard regression models assume that the dependent variable and regressors are sampled at the same frequency. However, in many practical applications—especially in *macroeconomics and finance*—important data sources are available at different frequencies:

- *Macroeconomic indicators* (e.g., GDP, inflation) are often reported quarterly or monthly.
- *Financial data* (e.g., stock prices, interest rates) are available at daily or intraday frequencies.

MIDAS models address this mismatch in data frequency by efficiently summarizing the impact of high-frequency data without requiring aggregation, interpolation, or data loss.

6.9.2 MIDAS Regression Specification

The MIDAS regression model extends the standard regression framework to allow for *mixed-frequency data*. A basic univariate MIDAS regression can be written as:

$$y_t = X_t'\beta + f\left(\left\{X_{t/S}^H\right\}, \theta, \lambda\right) + \epsilon_t. \tag{91}$$

where:

- y_t is the **low-frequency dependent variable**.
- $X_t'\beta$ represents the **low-frequency regressors**.
- $\left\{X_{t/S}^H\right\}$ represents a set of **high-frequency regressors**, where S high-frequency observations are available for each low-frequency time point.
- $f(\cdot)$ is a weighting function parameterized by θ and λ, which describes how the high-frequency data affects the low-frequency dependent variable.
- ϵ_t is the error term.

6.9.3 Key Features of MIDAS

- *No need for data aggregation*: Unlike traditional methods, MIDAS does not require averaging or summing high-frequency data, which can lead to information loss.
- *Parsimonious parameterization:* The function $f(\cdot)$ efficiently summarizes the impact of high-frequency regressors without overfitting.
- *Flexible weighting schemes*: MIDAS allows for different functional forms of $f(\cdot)$ to control how past high-frequency observations contribute to the low-frequency outcome.

6.9.4 MIDAS Weighting Functions

- **Step Weighting:**
  The step weighting approach assumes a fixed effect for all selected high-frequency regressors:

$$y_t = X_t'\beta + \sum_{\tau=0}^{k-1} \frac{X_{(t-\tau)}^H}{S}\varphi_\tau + \epsilon_t. \tag{92}$$

  This method applies an equal weight to each lagged high-frequency regressor within the aggregation window.

- **Almon Polynomial Weighting:**
  The Almon weighting scheme (also known as polynomial weighting) imposes a parametric polynomial structure on the weights, reducing the number of free parameters:

$$y_t = X_t'\beta + \sum_{\tau=0}^{k-1} \frac{X_{(t-\tau)}^H}{S}\left(\sum_{j=0}^{p} \tau^j\theta_j\right) + \epsilon_t. \tag{93}$$

  Here:

- The weights follow a polynomial structure with coefficients θ_j.
- The order of the polynomial p controls the flexibility of the weight function.

This approach allows MIDAS to capture time decay effects in a parsimonious manner.

- **Exponential Almon Weighting:**
  A refinement of the Almon polynomial approach, Exponential Almon Weighting, applies an exponential transformation to enforce smooth decay in the weighting function:

$$y_t = X_t'\beta + \sum_{\tau=0}^{k-1} \frac{X_{(t-\tau)}^H}{S} \left(\frac{\exp\left(\tau\theta_1 + \tau^2\theta_2\right)}{\sum_{j=0}^{k} \exp\left(j\theta_1 + j^2\theta_2\right)} \right) \lambda + \epsilon_t. \tag{94}$$

  This method ensures:

  - Smoothly decaying weights for older high-frequency observations.
  - More flexible functional forms than simple polynomials.

- **Beta Weighting:**
  The Beta weighting function generalizes the Exponential Almon approach by allowing more flexible non-linear decay:

$$y_t = X_t'\beta + \sum_{\tau=0}^{k-1} \frac{X_{(t-\tau)}^H}{S} \left(\frac{\exp\left(\tau\theta_1 + \tau^2\theta_2\right)}{\sum_{j=0}^{k} \exp\left(j\theta_1 + j^2\theta_2\right)} \right) \lambda + \epsilon_t. \tag{95}$$

  Beta weighting is particularly useful when the effect of high-frequency regressors varies non-monotonically over time.

6.9.5 The GARCH-MIDAS Model

The GARCH-MIDAS model is an extension of standard GARCH models, incorporating low-frequency macroeconomic variables into high-frequency financial volatility modeling. This allows for the decomposition of volatility into short-run and long-run components.

Motivation for GARCH-MIDAS:
Traditional GARCH models capture volatility clustering in financial time series but do not account for long-run variations in economic uncertainty. By integrating macroeconomic variables (e.g., inflation, interest rates, GDP) sampled at lower frequencies, GARCH-MIDAS models:

- Provide a better economic interpretation of volatility.
- Improve forecasting performance by utilizing long-term risk factors.

Model Specification:
The GARCH-MIDAS framework decomposes conditional variance into:

$$\sigma_{i,t}^2 = \tau_t \times g_{i,t}, \tag{96}$$

where:

- τ_t represents the long-run component (slow-moving, linked to macroeconomic variables).
- $g_{i,t}$ represents the short-run component (fast-moving, capturing financial volatility dynamics).

Short-Run Volatility Component The short-run component follows a standard GARCH(1,1) process:

$$g_{i,t} = (1 - \alpha - \beta) + \alpha \frac{(r_{i-1,t} - \mu)^2}{\tau_t} + \beta g_{i-1,t}, \tag{97}$$

where:

- $\alpha > 0, \beta \geq 0$, and $\alpha + \beta < 1$ ensure stationarity.
- Returns $r_{i,t}$ are modeled as:

$$r_{i,t} = \mu + \sqrt{\tau_t \times g_{i,t}} z_{i,t}, \quad z_{i,t} \sim N(0,1). \tag{98}$$

Long-Run Volatility Component The long-run component is modeled using a MIDAS polynomial filter applied to macroeconomic variables $X_{t,j}$:

$$\tau_t = \exp\left(m + \theta_j \sum_{k=1}^{K_j} \delta_{k,j}(\omega) X_{t-k,j} \right), \tag{99}$$

where:

- $\delta_{k,j}(\omega)$ is a weighting function (e.g., Almon or Beta weighting).
- θ_j controls the strength of macroeconomic effects.

Parameter Estimation The parameter set of the GARCH-MIDAS model is:

$$\Theta = \{\mu, \alpha, \beta, m, \theta_j, \omega_{1,j}, \omega_{2,j}\}, \quad j = 1, \ldots, J. \tag{100}$$

These parameters are estimated via Maximum Likelihood Estimation (MLE) by maximizing the log-likelihood function:

$$LLF = -\frac{1}{2} \sum_{t=1}^{T} \sum_{i=1}^{N_t} \left[\log(2\pi) + \log(g_{i,t}\tau_t) + \frac{(r_{i,t} - \mu)^2}{g_{i,t}\tau_t} \right]. \tag{101}$$

6.10 Multivariate GARCH (MGARCH) Models

Multivariate GARCH (MGARCH) models are used to forecast the conditional variances and covariances (or correlations) of multiple time series simultaneously. This is crucial in finance for applications like portfolio optimization, risk management, and asset pricing. We denote the $N \times 1$ vector of returns at time t as $r_t = \mu_t + \epsilon_t$, where $\mu_t = E[r_t|\mathcal{F}_{t-1}]$ is the conditional mean and ϵ_t is the vector of residuals (shocks) with $E[\epsilon_t|\mathcal{F}_{t-1}] = 0$. The conditional covariance matrix of ϵ_t (and thus r_t, assuming μ_t is known or estimated) is $H_t = E[\epsilon_t\epsilon_t'|\mathcal{F}_{t-1}]$.

6.10.1 Model Equations and Explanations

VECH Model The VECH model, proposed by Bollerslev, Engle, and Wooldridge (1988), directly models the elements of the conditional covariance matrix H_t. The 'vech()' operator stacks the lower (or upper) triangular portion of a symmetric $N \times N$ matrix H_t into an $N(N+1)/2 \times 1$ vector.

The general VECH(p,q) model is specified as:

$$\text{vech}(H_t) = C + \sum_{i=1}^{q} A_i \text{vech}(\epsilon_{t-i}\epsilon_{t-i}') + \sum_{j=1}^{p} B_j \text{vech}(H_{t-j}) \tag{102}$$

where:

- H_t is the $N \times N$ conditional covariance matrix at time t.
- ϵ_t is the $N \times 1$ vector of residuals at time t.
- C is an $N(N+1)/2 \times 1$ vector of constants.
- A_i and B_j are $N(N+1)/2 \times N(N+1)/2$ parameter matrices.

Meaning: Each conditional variance and covariance in H_t is a linear function of its own lagged values, lagged values of all other variances/covariances, past squared shocks, and past cross-products of shocks. **Pros:** Very general and flexible in describing the dynamics. **Cons:** The number of parameters ($N(N+1)/2$ for C, and $(p+q)(N(N+1)/2)^2$ for A_i, B_j) is extremely large, making estimation infeasible for even moderately sized N. Ensuring H_t is positive definite requires complex non-linear parameter restrictions.

Diagonal VECH (DVEC) Model The Diagonal VECH model simplifies the VECH model by restricting the parameter matrices A_i and B_j to be diagonal. The DVEC(p,q) model can be written element-wise for each $h_{kl,t}$ (the element in k-th row and l-th column of H_t):

$$h_{kl,t} = \omega_{kl} + \sum_{i=1}^{q} \alpha_{kli}\epsilon_{k,t-i}\epsilon_{l,t-i} + \sum_{j=1}^{p} \beta_{klj} h_{kl,t-j} \tag{103}$$

For the common DVEC(1,1) model, this simplifies to:

$$h_{kl,t} = \omega_{kl} + \alpha_{kl}\epsilon_{k,t-1}\epsilon_{l,t-1} + \beta_{kl}h_{kl,t-1} \tag{104}$$

where:

- $h_{kl,t}$ is the conditional covariance between asset k and asset l at time t (if $k = l$, it's the conditional variance of asset k).
- ω_{kl}, α_{kl}, β_{kl} are parameters specific to the (k, l)-th element.
- $\epsilon_{k,t-1}\epsilon_{l,t-1}$ is the product of lagged shocks for assets k and l.

Meaning: Each conditional variance or covariance term $h_{kl,t}$ follows its own GARCH-like process, depending only on its own lagged values and the lagged cross-product of the relevant residuals $\epsilon_{k,t-1}\epsilon_{l,t-1}$. There are no spillover effects from other distinct variance/covariance terms in the B_j terms (as B_j matrices are diagonal). **Pros:** Significantly reduces the number of parameters compared to the full VECH model (to $N(N+1)/2\times(1+p+q)$). **Cons:** While simpler, ensuring H_t is positive definite is still not guaranteed without further restrictions. The independent evolution of each covariance term might be too restrictive and not capture complex interactions well.

BEKK Model The BEKK (Baba, Engle, Kraft, and Kroner, 1995) model is designed to ensure the positive definiteness of the conditional covariance matrix H_t by its quadratic form. The general BEKK(p,q,K) model is:

$$H_t = C_0'C_0 + \sum_{k=1}^{K}\sum_{i=1}^{q} A_{ki}'\epsilon_{t-i}\epsilon_{t-i}'A_{ki} + \sum_{k=1}^{K}\sum_{j=1}^{p} B_{kj}'H_{t-j}B_{kj} \tag{105}$$

A common specification is the BEKK(1,1) model (setting $K = 1$, $p = 1$, $q = 1$):

$$H_t = C_0'C_0 + A'\epsilon_{t-1}\epsilon_{t-1}'A + B'H_{t-1}B \tag{106}$$

where:

- C_0 is an $N \times N$ lower (or upper) triangular matrix of constants. $C_0'C_0$ ensures the intercept matrix is positive semi-definite.
- A and B are $N \times N$ parameter matrices.
- $\epsilon_{t-1}\epsilon_{t-1}'$ is the outer product of lagged residual vectors.
- H_{t-1} is the lagged conditional covariance matrix.

Diagonal BEKK Model (1,1): A simplification where A and B are diagonal matrices. If we denote these diagonal matrices as A_1 and B_1, and the constant term as $M = C_0'C_0$ (or more generally, a symmetric positive semi-definite matrix):

$$H_t = M + A_1\epsilon_{t-1}\epsilon_{t-1}'A_1' + B_1H_{t-1}B_1' \tag{107}$$

(Note: If A_1, B_1 are diagonal, $A_1' = A_1, B_1' = B_1$. The terms are $A_1\epsilon_{t-1}\epsilon_{t-1}'A_1$ and $B_1H_{t-1}B_1$). An extension to include asymmetric effects (leverage) as seen in some software outputs can be:

$$H_t = M + A_1\epsilon_{t-1}\epsilon_{t-1}'A_1' + D_1\xi_{t-1}\xi_{t-1}'D_1' + B_1H_{t-1}B_1' \tag{108}$$

where:

- M is an $N \times N$ symmetric positive semi-definite intercept matrix.
- A_1, D_1, B_1 are $N \times N$ diagonal parameter matrices.
- ϵ_{t-1} is the vector of lagged residuals.
- ξ_{t-1} is a vector where its i-th element $\xi_{i,t-1} = \epsilon_{i,t-1}$ if $\epsilon_{i,t-1} < 0$ and 0 otherwise, capturing asymmetric responses to negative shocks.

Meaning: The BEKK model specifies H_t such that it is positive definite by construction (if M is PSD). The conditional covariances are influenced by past shocks and past covariances through matrix multiplications, allowing for interactions. The diagonal version restricts these interactions but is more parsimonious. The asymmetric term D_1 allows negative shocks to have a different impact on volatility/covariance than positive shocks. **Pros:** Guarantees H_t is positive definite. More parsimonious than VECH, especially the diagonal BEKK. **Cons:** Full BEKK can still have many parameters ($N(N+1)/2$ for $C_0'C_0$, plus $2KN^2$ for A_{ki}, B_{kj} in the general summation form). Estimation can be complex. The interpretation of individual parameters in A and B (if not diagonal) can be difficult.

Constant Conditional Correlation (CCC) Model The Constant Conditional Correlation (CCC) model, proposed by Bollerslev (1990), simplifies modeling by assuming that conditional correlations are constant over time. The conditional covariance matrix H_t is specified as:

$$H_t = D_tRD_t \tag{109}$$

where:

- $D_t = \text{diag}(\sqrt{h_{11,t}}, \sqrt{h_{22,t}}, ..., \sqrt{h_{NN,t}})$ is an $N \times N$ diagonal matrix of time-varying conditional standard deviations.
- Each conditional variance $h_{ii,t}$ (for $i = 1, ..., N$) is typically modeled by a univariate GARCH(p,q) process:

$$h_{ii,t} = \omega_i + \sum_{k=1}^{q_i} \alpha_{ik}\epsilon_{i,t-k}^2 + \sum_{l=1}^{p_i} \beta_{il}h_{ii,t-l} \tag{110}$$

- R is the $N \times N$ time-invariant (constant) conditional correlation matrix. R is symmetric, with $R_{ii} = 1$ and $R_{ij} = \rho_{ij}$ for $i \neq j$. R must be positive definite.

The conditional covariance between asset i and asset j is $h_{ij,t} = \rho_{ij}\sqrt{h_{ii,t}h_{jj,t}}$. **Meaning:** The CCC model separates the modeling of individual conditional variances from the modeling of conditional correlations. Variances are time-varying according to univariate GARCH processes, but their correlations are assumed to be fixed. **Pros:**

- Computationally simple; can be estimated in two stages (univariate GARCH models first, then R from standardized residuals).
- Guarantees H_t is positive definite if R is positive definite and all $h_{ii,t} > 0$.
- Parsimonious: $N\times(1+p+q)$ parameters for variances (if all GARCH(p,q)) plus $N(N-1)/2$ correlation parameters.

Cons: The assumption of constant correlations is often unrealistic, especially in turbulent market periods where correlations are known to change.

Dynamic Conditional Correlation (DCC) Model The Dynamic Conditional Correlation (DCC) model, developed by Engle (2002), extends the CCC model by allowing conditional correlations to vary over time. The conditional covariance matrix H_t is specified as:

$$H_t = D_t R_t D_t \tag{111}$$

where:

- D_t is the same diagonal matrix of conditional standard deviations as in the CCC model, with $h_{ii,t}$ typically following univariate GARCH processes (Equation 110).
- R_t is the $N \times N$ time-varying conditional correlation matrix.

R_t is constructed from a pseudo-correlation matrix Q_t:

$$Q_t = (1-\alpha-\beta)\bar{Q} + \alpha u_{t-1}u'_{t-1} + \beta Q_{t-1} \tag{112}$$

And R_t is obtained by standardizing Q_t:

$$R_t = (\text{diag}(Q_t))^{-1/2} Q_t (\text{diag}(Q_t))^{-1/2} \tag{113}$$

where:

- $u_t = D_t^{-1}\epsilon_t$ is the vector of standardized residuals, i.e., $u_{i,t} = \epsilon_{i,t}/\sqrt{h_{ii,t}}$.
- $\bar{Q}$ is the $N \times N$ unconditional covariance matrix of the standardized residuals u_t. It is typically estimated as the sample covariance matrix of u_t.
- α and β are non-negative scalar parameters governing the dynamics of Q_t. For Q_t to be stationary and positive definite, $\alpha + \beta < 1$, $\alpha \geq 0$, $\beta \geq 0$. Q_0 must be positive definite.

Meaning: The DCC model allows correlations to evolve over time based on past standardized shocks and past correlation structures (via Q_t). The GARCH-like dynamics for Q_t mean that correlations can exhibit persistence and react to market events. **Pros:**

- More realistic than CCC by allowing time-varying correlations.
- Still relatively parsimonious (only two additional parameters, α and β, for the correlation dynamics beyond the univariate GARCH models and $\bar{Q}$).
- Estimation can be performed in stages: 1. Univariate GARCH models. 2. $\bar{Q}$. 3. α, β.
- H_t is positive definite if Q_t is positive definite (which is ensured if $(1 - \alpha - \beta)\bar{Q}$ is PSD and Q_0 is PSD).

Cons:

- The dynamic structure imposed by scalar α and β is common to all correlations, which might be restrictive. All correlations respond with the same speed to shocks.
- Two-step estimation is consistent but not fully efficient. Full MLE is more complex.

6.10.2 Challenges in Multivariate Volatility Forecasting

Forecasting multivariate volatility presents several significant challenges:

1. **Curse of Dimensionality:** As the number of assets (N) increases, the number of parameters in MGARCH models can explode, especially for general models like VECH or full BEKK. This makes estimation difficult, imprecise, or even infeasible.
    - *Addressing this:* Diagonal VECH, Diagonal BEKK, CCC, and DCC models are more parsimonious and aim to reduce parameter count. For example, DCC only adds two parameters (α, β) for correlation dynamics on top of univariate GARCH models.
2. **Positive Definiteness:** The conditional covariance matrix H_t must be positive definite (or at least positive semi-definite) at each point in time to be valid.
    - *Addressing this:* BEKK models ensure positive definiteness by their quadratic structure. CCC and DCC models ensure it if the correlation matrix (R or R_t) is positive definite and univariate variances are positive. VECH and DVEC models require complex parameter restrictions to guarantee this.

3. **Capturing Dynamic Correlations:** Correlations between asset returns are not constant; they change over time, often increasing during market stress.

    - *Addressing this:* CCC model assumes constant correlation, which is a limitation. VECH, BEKK, and DCC models allow for time-varying correlations. DCC is specifically designed to model these dynamics in a parsimonious way. BEKK offers rich dynamics but can be parameter-heavy unless restricted (e.g., diagonal BEKK).

4. **Asymmetric Effects (Leverage):** Volatility often responds more to negative shocks than to positive shocks of the same magnitude.

    - *Addressing this:* This can be incorporated by using asymmetric GARCH models (e.g., GJR-GARCH) for the univariate variances in CCC/DCC, or by adding specific asymmetric terms in VECH or BEKK models (as in Equation 108).

5. **Estimation Complexity:** Maximizing the likelihood function for MGARCH models can be computationally intensive due to the large number of parameters and the non-linearity of the models.

    - *Addressing this:* CCC and DCC models often use multi-stage estimation procedures which are computationally less demanding, though potentially less efficient than full information maximum likelihood (FIML). Parsimonious models like Diagonal BEKK are also easier to estimate than their full counterparts.

6. **Model Specification and Selection:** Choosing the appropriate MGARCH model, lag orders (p,q), and distributional assumptions for residuals is a complex task.

6.10.3 Significance and Flexibility of Correlations in MGARCH Models

Conditional correlations are a key output of MGARCH models, vital for financial decision-making. The ability of a model to capture correlation dynamics adequately is a critical aspect of its usefulness.

- **VECH Model:**
    - *Significance:* The full VECH models individual covariances directly, so their significance can be tested via the parameters in C, A_i, B_j.
    - *Flexibility:* Extremely flexible in theory, as $h_{ij,t}$ can depend on all past shocks and covariances. However, this leads to overparameterization. Practical, restricted versions are needed, which reduces flexibility.
- **Diagonal VECH (DVEC) Model:**

- *Significance:* The parameters $\omega_{ij}, \alpha_{ij}, \beta_{ij}$ for each $h_{ij,t}$ determine its dynamics. Significance of these indicates how past shocks/covariances influence current covariance.
- *Flexibility:* Correlations $\rho_{ij,t} = h_{ij,t}/\sqrt{h_{ii,t}h_{jj,t}}$ are time-varying. However, each $h_{ij,t}$ evolves independently based only on its own lags and $\epsilon_{i,t-1}\epsilon_{j,t-1}$. This limits the model's ability to capture common factors driving correlation changes across many assets. It might not be flexible enough for complex correlation structures.

- **BEKK Model (e.g., Diagonal BEKK with asymmetry):**
  - *Significance:* The significance of off-diagonal elements in the intercept matrix M (e.g., M_{ij} for $i \neq j$) directly indicates significant constant components of covariance. The significance of elements in (diagonal) A_1, D_1, B_1 indicates how past shocks, asymmetric shocks, and past covariances influence current covariances. For example, in a bivariate diagonal BEKK, $h_{12,t} = M_{12} + A_{1,11}A_{1,22}\epsilon_{1,t-1}\epsilon_{2,t-1} + D_{1,11}D_{1,22}\xi_{1,t-1}\xi_{2,t-1} + B_{1,11}B_{1,22}h_{12,t-1}$ (where $A_{1,ii}$ is the i-th diagonal element of A_1).
  - *Flexibility:* Full BEKK is very flexible. Diagonal BEKK allows for time-varying correlations. The interaction terms ensure that correlations are dynamic. However, the diagonal structure restricts how shocks to one asset influence the covariances involving other assets compared to a full BEKK. It is more flexible than CCC or DVEC in some respects due to the multiplicative structure but less flexible than full BEKK or potentially DCC for correlation-specific dynamics. Whether it is "flexible enough" depends on the empirical application; it may capture broad co-movement but perhaps not subtle changes in correlation structure.

- **Constant Conditional Correlation (CCC) Model:**
  - *Significance:* The constant correlation coefficients ρ_{ij} in matrix R are estimated. Their statistical significance indicates a stable, average relationship.
  - *Flexibility:* By definition, this model is *not flexible* for forecasting correlations, as it assumes they are constant. It cannot capture changes in correlation structure over time. If correlations are indeed stable, it is parsimonious and efficient. However, for assets like energy futures where correlations can shift due to market conditions, supply shocks, or geopolitical events, CCC is likely too restrictive.

- **Dynamic Conditional Correlation (DCC) Model:**
  - *Significance:* The parameters α and β determine the dynamic behavior of correlations. If α and β are significant (and $\alpha + \beta > 0$), it indicates that correlations are indeed time-varying. The elements of $\bar{Q}$ represent the long-run target correlations.

– *Flexibility:* The DCC model is designed to be flexible in capturing time-varying correlations. The GARCH-like dynamics for Q_t allow correlations to be persistent and to react to market-wide shocks (as captured by standardized residuals). It is generally considered flexible enough for many applications, providing a good balance between parsimony and dynamic richness. However, the scalar α, β imply common dynamics for all correlations, which could be a limitation if different pairs of assets have very different correlation behaviors. More advanced DCC variants (e.g., ADCC, cDCC) offer further flexibility.

In summary, when forecasting correlations, especially for volatile assets like energy futures, models that allow for dynamic correlations (BEKK, DCC) are generally preferred over models assuming constant correlations (CCC). The choice between BEKK and DCC (or other dynamic models) often depends on the number of assets, desired parsimony, and specific characteristics of the data. Diagonal BEKK can be a good compromise for tractability, while DCC offers intuitive modeling of correlation dynamics.

6.11 Volatility Forecasting and Evaluation

6.11.1 Multi-Step Volatility Forecasting

Alternative Approaches:
Multi-period forecasting can be conducted using different methodologies. The three primary approaches are:

- **Scaled forecasts**
- **Iterated forecasts**
- **Direct forecasts**

Scaled Forecasts:
A scaled forecast aims to predict the variance of returns h steps ahead (r_{t+h}) using a scaling relationship. The variance expectation follows:

$$\mathbb{E}(\epsilon_{t+h}^2 \mid \mathcal{F}_t) = \mathbb{E}(\sigma_{t+1}^2 \mid \mathcal{F}_t) \times \sqrt{h} \tag{114}$$

where $\mathcal{F}_t$ represents the information set available at time t.

Direct Forecasts:
A direct forecast estimates the variance of future returns for a given horizon h directly, assuming that the expectation of future shocks is zero. The formulation is given by:

$$\mathbb{E}(\epsilon_{t+h}^2 \mid \mathcal{F}_t) = \mathbb{E}(\sigma_{t+h}^2 \mid \mathcal{F}_t) \tag{115}$$

where the variance is estimated independently for each forecast horizon.

Iterated Forecasts:
In an iterated forecast approach, the multi-period variance is computed recursively using an autoregressive specification. The one-step-ahead variance forecast follows:

$$\mathbb{E}(\sigma_{t+1}^2 \mid \mathcal{F}_t) = \alpha_0 + \alpha_1 \epsilon_t^2 + \beta_1 \sigma_t^2 \tag{116}$$

Subsequent steps are calculated iteratively:

$$\mathbb{E}(\sigma_{t+h}^2 \mid \mathcal{F}_t) = \alpha_0 + (\alpha_1 + \beta_1)\mathbb{E}(\sigma_{t+h-1}^2 \mid \mathcal{F}_t) \tag{117}$$

Empirical Performance:
Empirical studies on multi-period volatility forecasting suggest:

- Scaled forecasts generally do not perform well.
- There are mixed results when comparing direct vs. iterated forecasts vs. MIDAS methods.

6.11.2 Volatility Forecast Evaluation

Mean Squared Error (MSE):
One common metric for evaluating volatility forecasts is the Mean Squared Error (MSE), which measures the average squared deviation between the forecast and the realized volatility. It is defined as:

$$MSE(RV_t, F_t) \equiv (RV_t - F_t)^2, \tag{118}$$

where F_t represents the forecasted volatility and RV_t represents the realized volatility.

QLIKE Loss Function:
The QLIKE (Quasi-Likelihood) loss function is another popular metric, particularly robust to noise in volatility estimates. It is defined as:

$$QLIKE(RV_t, F_t) \equiv \frac{RV_t}{F_t} - \ln\left(\frac{RV_t}{F_t}\right) - 1. \tag{119}$$

QLIKE is particularly useful because it penalizes large underpredictions more heavily than small overpredictions, making it suitable for financial applications where volatility misestimation can have asymmetric consequences.

7 Guide for Modeling Financial time series: ARMA-GARCH (Not in book)

Modeling financial time series often involves capturing both the conditional mean (average behavior) and the conditional variance (volatility). The ARMA-GARCH framework is a popular approach for this. This guide outlines the general steps involved.

Step 1: Data Preparation and Preliminary Analysis

1. **Obtain and Clean Data:** Collect the raw price series (P_t) for the financial asset of interest. Ensure data quality (e.g., handle missing values, outliers if appropriate, though GARCH models are designed to handle some outlier-like behavior due to volatility).

2. **Calculate Returns:** Financial time series analysis is typically performed on returns, not prices, as returns are more likely to be stationary. Log returns are often preferred:

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

   Why? Log returns have convenient properties (e.g., additivity over time, approximate percentage change for small returns) and often exhibit better statistical properties (closer to normality, more stationary) than simple returns.

3. **Visual Inspection of Returns:** Plot the time series of returns. **Why?** To observe key stylized facts of financial returns:

   - **Stationarity (Mean Reversion):** Does the series appear to fluctuate around a constant mean?
   - **Volatility Clustering:** Are there periods of high volatility followed by periods of high volatility, and periods of low volatility followed by periods of low volatility? This is a key indicator for GARCH models.
   - **Spikes/Fat Tails:** Are there extreme observations (leptokurtosis)?

Step 2: Test for Stationarity of Returns

Before modeling, formally test if the return series (r_t) is stationary.

1. **Formal Tests:** Use tests like the Augmented Dickey-Fuller (ADF) test (null hypothesis: unit root, i.e., non-stationary) or the Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test (null hypothesis: stationarity). **Why?** Most time series models, including ARMA, assume stationarity. If returns are non-stationary (uncommon for daily or higher frequency returns), differencing might be required, though this is usually addressed by using returns instead of prices.

Step 3: Modeling the Conditional Mean (ARMA(p,q) Process)

The goal is to capture any linear serial dependence in the returns.

$$r_t = c + \sum_{i=1}^{p} \phi_i r_{t-i} + \epsilon_t + \sum_{j=1}^{q} \theta_j \epsilon_{t-j}$$

where ϵ_t are the residuals (innovations) from the mean equation.

1. **Identify Potential ARMA Orders (p,q):** Examine the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots of the return series. **Why?**

   - For an AR(p) process, the PACF cuts off after lag p.
   - For an MA(q) process, the ACF cuts off after lag q.
   - For an ARMA(p,q) process, both ACF and PACF decay gradually.

   Often, for financial returns, p and q are small or even zero (i.e., returns are close to a white noise process or only have a constant mean).

2. **Estimate Candidate ARMA Models:** Fit several plausible ARMA(p,q) models based on ACF/PACF insights. Include a constant term (c) unless there's a strong reason not to.

3. **Select the Best ARMA Model:** Use information criteria like AIC (Akaike Information Criterion) or BIC (Bayesian Information Criterion) to select the most parsimonious model that fits well. Also, check the statistical significance of the estimated coefficients. **Why?** To balance model fit with model complexity, avoiding overfitting.

4. **Diagnostic Checking of ARMA Residuals (ϵ_t):** Once an ARMA model is chosen, examine its residuals:

   - **Plot residuals:** They should appear to be random, centered around zero.
   - **ACF/PACF of residuals:** There should be no significant spikes, indicating that the ARMA model has captured the linear dependence.
   - **Ljung-Box Q-test on residuals:** Formally tests for remaining serial correlation in the residuals (null hypothesis: no serial correlation).

   Why? The residuals from a well-specified mean model should be white noise (uncorrelated). If not, the mean model needs refinement.

Step 4: Test for ARCH Effects in ARMA Residuals

After fitting the mean model, test if the residuals (ϵ_t) exhibit time-varying volatility (ARCH effects).

1. **Visual Inspection of Squared Residuals:** Plot the squared residuals (ϵ_t^2). Look for clustering patterns similar to those observed in the original returns plot. **Why?** ϵ_t^2 is a proxy for the variance. Clustering in squared residuals suggests that volatility is autocorrelated.

2. **ACF/PACF of Squared Residuals:** Examine the ACF and PACF of ϵ_t^2. Significant spikes indicate serial correlation in the variance. **Why?** This is a common way to detect ARCH effects.

3. **Formal Tests for ARCH Effects:**

   - **Ljung-Box Q-test on squared residuals (ϵ_t^2):** Tests for serial correlation in the squared residuals.
   - **Engle's ARCH-LM (Lagrange Multiplier) test:** Specifically tests for ARCH effects (null hypothesis: no ARCH effects). This involves regressing squared residuals on their lagged values.

   Why? Formal tests provide statistical evidence for or against the presence of ARCH effects. If ARCH effects are present, a GARCH model for the conditional variance is appropriate.

Step 5: Modeling the Conditional Variance (GARCH(P,Q) Process)

If ARCH effects are detected, model the conditional variance $\sigma_t^2 = \text{Var}(r_t|\mathcal{F}_{t-1}) = \text{Var}(\epsilon_t|\mathcal{F}_{t-1})$. A common GARCH(P,Q) model is:

$$\sigma_t^2 = \omega + \sum_{i=1}^{P} \alpha_i \epsilon_{t-i}^2 + \sum_{j=1}^{Q} \beta_j \sigma_{t-j}^2$$

where $\omega > 0$, $\alpha_i \geq 0$, $\beta_j \geq 0$ are parameters.

1. **Specify and Estimate the Combined ARMA-GARCH Model:** Simultaneously estimate the parameters of the chosen ARMA(p,q) model for the mean and a GARCH(P,Q) model for the variance. GARCH(1,1) is often a good starting point. **Why?** The mean and variance equations are typically estimated together using Maximum Likelihood Estimation (MLE), assuming a distribution for the standardized innovations $z_t = \epsilon_t/\sigma_t$ (e.g., Normal, Student's t).

2. **Select GARCH Orders (P,Q):** If GARCH(1,1) is insufficient (based on diagnostics), try other low-order models (e.g., GARCH(1,2), GARCH(2,1)). Use information criteria (AIC, BIC) and parameter significance to guide selection. **Why?** Parsimony is important. Higher-order GARCH models can be difficult to estimate and may not offer significant improvements.

3. **Check Parameter Constraints and Significance:**

   - Ensure $\omega > 0$, $\alpha_i \geq 0$, $\beta_j \geq 0$.
   - For covariance stationarity of the GARCH process, typically $\sum \alpha_i + \sum \beta_j < 1$.
   - Check if estimated parameters are statistically significant.

   Why? These conditions are necessary for a well-behaved and interpretable GARCH model.

Step 6: Diagnostic Checking of the Full ARMA-GARCH Model

After estimating the ARMA-GARCH model, assess its adequacy by examining the standardized residuals: $z_t = \epsilon_t / \hat{\sigma}_t$.

1. **Examine Standardized Residuals (z_t):**
   - **Plot z_t:** Should look like an i.i.d. white noise series, ideally matching the assumed distribution (e.g., Normal, t).
   - **ACF/PACF of z_t:** Should show no significant serial correlation.
   - **Ljung-Box Q-test on z_t:** Formally test for serial correlation.

   Why? The standardized residuals should be free of any remaining linear dependence not captured by the mean equation.

2. **Examine Squared Standardized Residuals (z_t^2):**
   - **Plot z_t^2:** Should show no remaining volatility clustering.
   - **ACF/PACF of z_t^2:** Should show no significant serial correlation.
   - **Ljung-Box Q-test on z_t^2:** Formally test for remaining serial correlation in variance.
   - **ARCH-LM test on z_t:** Test if any ARCH effects remain.

   Why? A well-specified GARCH model should capture all the conditional heteroskedasticity, leaving the squared standardized residuals free of ARCH effects.

3. **Goodness-of-Fit Tests:** If a specific distribution (e.g., Student's t) was assumed for z_t, perform goodness-of-fit tests (e.g., Kolmogorov-Smirnov, Jarque-Bera if normality was assumed) to check if the standardized residuals conform to this distribution. **Why?** The choice of distribution for the innovations affects parameter estimates and subsequent applications like VaR estimation.

If diagnostics indicate issues, revisit model specification (e.g., ARMA orders, GARCH orders, assumed distribution, or consider alternative GARCH variants like EGARCH, GJR-GARCH if asymmetry/leverage effects are suspected).

Step 7: Interpretation and Application

Once a satisfactory model is found:

1. **Interpret Parameters:** Understand the implications of the estimated coefficients (e.g., persistence of volatility from $\alpha_1 + \beta_1$ in a GARCH(1,1)).

2. **Forecasting:** Use the model to forecast future conditional mean and, more importantly, future conditional volatility.

3. **Risk Management:** Apply the model for Value-at-Risk (VaR) calculations, option pricing, portfolio optimization, etc.

This systematic approach helps in building robust models for financial time series that can capture their key dynamic features.

8 Multivariate Models

8.1 Vector Autoregressive Models (VARs)

Vector autoregressive models are a natural generalization of univariate autoregressive models. A VAR is a systems regression model (i.e., there is more than one dependent variable). The simplest case is a bivariate VAR where there are only two variables y_{1t} and y_{2t}, each of whose current values depend on different combinations of the previous k values of both variables, and error terms.

$$y_{1t} = \beta_{10} + \beta_{11} y_{1t-1} + \cdots + \beta_{1k} y_{1t-k} + \alpha_{11} y_{2t-1} + \cdots + \alpha_{1k} y_{2t-k} + u_{1t} \quad (120)$$

$$y_{2t} = \beta_{20} + \beta_{21} y_{2t-1} + \cdots + \beta_{2k} y_{2t-k} + \alpha_{21} y_{1t-1} + \cdots + \alpha_{2k} y_{1t-k} + u_{2t} \quad (121)$$

where u_{it} is a white noise disturbance term with $E(u_{it}) = 0$, $(i = 1, 2)$.

Although, for simplicity, we usually assume that $E(u_{1t}u_{2t}) = 0$ so that the disturbances are uncorrelated across equations, it is common and more realistic to allow them to be contemporaneously correlated, so $\text{Cov}(u_{1t}, u_{2t}) = \sigma_{12}$.

As should already be evident, an important feature of the VAR model is its flexibility and the ease of generalization. For example, the model could be extended to encompass moving average errors, which would be a multivariate version of an ARMA model, known as a VARMA. Instead of having only two variables, y_{1t} and y_{2t}, the system could also be expanded to include g variables, $y_{1t}, y_{2t}, y_{3t}, \ldots, y_{gt}$, each of which has an equation.

Another useful facet of VAR models is the compactness with which the notation can be expressed. For example, consider the case from above where $k = 1$, so that each variable depends only upon the immediately previous values of y_{1t} and y_{2t}, plus an error term. This could be written as:

$$y_{1t} = \beta_{10} + \beta_{11} y_{1t-1} + \alpha_{11} y_{2t-1} + u_{1t} \quad (122)$$

$$y_{2t} = \beta_{20} + \beta_{21} y_{2t-1} + \alpha_{21} y_{1t-1} + u_{2t} \quad (123)$$

or

$$\begin{pmatrix} y_{1t} \\ y_{2t} \end{pmatrix} = \begin{pmatrix} \beta_{10} \\ \beta_{20} \end{pmatrix} + \begin{pmatrix} \beta_{11} & \alpha_{11} \\ \alpha_{21} & \beta_{21} \end{pmatrix} \begin{pmatrix} y_{1t-1} \\ y_{2t-1} \end{pmatrix} + \begin{pmatrix} u_{1t} \\ u_{2t} \end{pmatrix} \quad (124)$$

or even more compactly as

$$y_t = \underbrace{\beta_0}_{g\times 1} + \underbrace{\beta_1}_{g\times g}\underbrace{y_{t-1}}_{g\times 1} + \underbrace{u_t}_{g\times 1} \tag{125}$$

In the previous equation there are $g = 2$ variables in the system. Extending the model to the case where there are k lags of each variable in each equation is also easily accomplished using this notation:

$$y_t = \underbrace{\beta_0}_{g\times 1} + \underbrace{\beta_1}_{g\times g}\underbrace{y_{t-1}}_{g\times 1} + \underbrace{\beta_2}_{g\times g}\underbrace{y_{t-2}}_{g\times 1} + \cdots + \underbrace{\beta_k}_{g\times g}\underbrace{y_{t-k}}_{g\times 1} + \underbrace{u_t}_{g\times 1} \tag{126}$$

8.1.1 Advantages and Disadvantages in VAR modeling

Advantages:

- **Endogeneity Handling:** Unlike structural models that require identifying exogenous and endogenous variables, VAR treats all variables as endogenous, avoiding restrictive identification assumptions.
- **Flexibility:** VARs extend beyond univariate AR models, allowing a variable to depend on its own lags as well as those of other variables, capturing richer dynamic structures.
- **Ease of Estimation:** When contemporaneous terms are absent on the right-hand side (RHS) and disturbances are uncorrelated, each equation can be estimated separately using ordinary least squares (OLS). In cases of correlated disturbances, maximum likelihood estimation (MLE) can be used.
- **Forecasting Performance:** Studies suggest that VAR models often outperform large-scale structural models in out-of-sample forecasting, particularly for macroeconomic variables such as GDP and unemployment rates.

Disadvantages:

- **Lack of Theoretical Structure:** VARs are largely empirical and do not incorporate strong theoretical restrictions, making policy interpretation more difficult.
- **Parameter Proliferation:** With g variables and k lags per equation, VAR models require estimating $g + kg^2$ parameters, leading to high-dimensional estimation problems, particularly with limited sample sizes.
- **Lag Length Selection:** Determining the appropriate number of lags is nontrivial, requiring information criteria (e.g., AIC, BIC) or other heuristics.
- **Stationarity Concerns:** Standard hypothesis tests require stationary data, but some proponents argue against differencing, as it may remove long-run relationships. This trade-off complicates model specification and interpretation.

8.1.2 Choosing the Optimal Lag Length for a VAR

Often, financial theory will have little to say on what is an appropriate lag length for a VAR and how long changes in the variables should take to work through the system. In such instances, there are broadly three methods that could be used to arrive at the optimal lag length: rules of thumb, cross-equation restrictions and information criteria.

Rules of Thumb for VAR Lag Length Selection:
Similar to univariate AR(p) models, it might be possible to use the data frequency to decide the lag order and thus, for example, selecting $p = 5$ for daily data, $p = 4$ for quarterly data, and so on. However, if the number of variables in the system is quite large, then a value of p this big, let alone the number that would by analogy be suggested for monthly data, would quickly become infeasible. It is also common to use an arbitrary fixed number of lags (typically, 1, 2, or 3) without further testing.

It is worth noting that a high value of p may be required if the number of variables g in the system is too small and excludes relevant influences on the included variables (possibly also including the wrong variables). Thus, in a sense, there is a trade-off between a larger p and a larger g. In such a situation, a better model is likely to arise from thinking creatively about additional variables to include in the model, even if these are hard to estimate, rather than increasing the lag length.

Cross-Equation Restrictions for VAR Lag Length Selection:
Determining the appropriate lag length in a VAR model requires testing the joint significance of all lags across all equations, rather than performing separate F-tests for each equation. A VAR with different lag lengths for different equations is effectively a restricted VAR, as some coefficients are implicitly set to zero.

A common approach is to specify the same lag order for all equations and use a likelihood ratio (LR) test to determine whether longer lags contribute significantly to the model. Suppose a quarterly VAR is estimated with eight lags per variable, and we wish to test whether lags five to eight can be excluded. The LR test statistic is given by:

$$LR = T \ln \left(\frac{|\Sigma_r|}{|\Sigma_u|} \right) \tag{127}$$

where:

- $|\Sigma_r|$ is the determinant of the variance-covariance matrix of residuals from the restricted model (with fewer lags).
- $|\Sigma_u|$ is the determinant of the variance-covariance matrix from the unrestricted model.
- T is the sample size.

The test statistic follows an asymptotic χ^2 distribution with degrees of freedom equal to the total number of restrictions. If the restriction imposes q lag exclusions in a VAR with g equations, the number of restrictions is:

$$g^2 q \tag{128}$$

This test assesses whether the restricted model leads to a significant increase in residual sum of squares (RSS). If $|\Sigma_r|$ and $|\Sigma_u|$ are similar, the restriction is supported, indicating that the additional lags are unnecessary.

Information Criteria for VAR Lag Length Selection:

While the likelihood ratio (LR) test is a straightforward method for selecting lag length in VAR models, it has notable limitations. First, the test can only compare two models at a time, where one must be a restricted version of the other. This makes it difficult to determine the optimal lag length directly. For instance, if the best lag length were seven or ten, testing only nested models (e.g., comparing VAR(8) to VAR(4)) might not reveal this.

Additionally, the LR test relies on the assumption that residuals are normally distributed, which is often unrealistic for financial data. As an alternative, information criteria—commonly used in ARMA model selection—offer a more flexible approach, as they do not assume normality. These criteria balance model fit and complexity by penalizing the inclusion of additional lags.

Multivariate versions of commonly used information criteria include:

$$MAIC = \ln|\hat{\Sigma}| + \frac{2k'}{T} \tag{129}$$

$$MSBIC = \ln|\hat{\Sigma}| + \frac{k'}{T}\ln(T) \tag{130}$$

$$MHQIC = \ln|\hat{\Sigma}| + \frac{2k'}{T}\ln(\ln(T)) \tag{131}$$

where:

- $\hat{\Sigma}$ is the variance-covariance matrix of residuals,
- T is the number of observations,
- k' is the total number of regressors in all equations.

For a VAR system with p equations, each containing k lags of p variables, plus a constant term, the number of estimated parameters is given by:

$$p^2 k + p \tag{132}$$

To determine the optimal lag length, models are estimated for a range of lags (e.g., from 0 to a pre-specified maximum $\bar{k}$), and the lag length minimizing the chosen information criterion is selected.

8.1.3 Does the VAR Include Contemporaneous Terms?

VAR models are typically specified without contemporaneous terms, meaning that all variables on the right-hand side (RHS) are either lagged values of the dependent variables or exogenous variables. This ensures that the model can be estimated equation by equation using ordinary least squares (OLS). The standard VAR form is:

$$y_{1t} = \beta_{10} + \beta_{11} y_{1t-1} + \alpha_{11} y_{2t-1} + u_{1t} \tag{133}$$
$$y_{2t} = \beta_{20} + \beta_{21} y_{2t-1} + \alpha_{21} y_{1t-1} + u_{2t} \tag{134}$$

However, if contemporaneous dependencies exist between variables, the model includes feedback terms, leading to a *VAR in primitive form*:

$$y_{1t} = \beta_{10} + \beta_{11} y_{1t-1} + \alpha_{11} y_{2t-1} + \alpha_{12} y_{2t} + u_{1t} \tag{135}$$
$$y_{2t} = \beta_{20} + \beta_{21} y_{2t-1} + \alpha_{21} y_{1t-1} + \alpha_{22} y_{1t} + u_{2t} \tag{136}$$

Stacking the system into matrix form:

$$\begin{pmatrix} y_{1t} \\ y_{2t} \end{pmatrix} = \begin{pmatrix} \beta_{10} \\ \beta_{20} \end{pmatrix} + \begin{pmatrix} \beta_{11} & \alpha_{11} \\ \alpha_{21} & \beta_{21} \end{pmatrix} \begin{pmatrix} y_{1t-1} \\ y_{2t-1} \end{pmatrix} + \begin{pmatrix} \alpha_{12} & 0 \\ 0 & \alpha_{22} \end{pmatrix} \begin{pmatrix} y_{2t} \\ y_{1t} \end{pmatrix} + \begin{pmatrix} u_{1t} \\ u_{2t} \end{pmatrix} \tag{137}$$

A *structural VAR (primitive form)* contains contemporaneous terms, resembling a system of simultaneous equations. However, this form is not directly identifiable, as the same predetermined (lagged) variables appear on both RHS equations. To achieve identification, one of the contemporaneous coefficients (α_{12} or α_{22}) must be set to zero.

By moving the contemporaneous terms to the left-hand side:

$$\begin{pmatrix} 1 & -\alpha_{12} \\ -\alpha_{22} & 1 \end{pmatrix} \begin{pmatrix} y_{1t} \\ y_{2t} \end{pmatrix} = \begin{pmatrix} \beta_{10} \\ \beta_{20} \end{pmatrix} + \begin{pmatrix} \beta_{11} & \alpha_{11} \\ \alpha_{21} & \beta_{21} \end{pmatrix} \begin{pmatrix} y_{1t-1} \\ y_{2t-1} \end{pmatrix} + \begin{pmatrix} u_{1t} \\ u_{2t} \end{pmatrix} \tag{138}$$

or in compact notation:

$$A y_t = \beta_0 + \beta_1 y_{t-1} + u_t \tag{139}$$

Pre-multiplying by A^{-1}, we obtain the *standard form VAR*:

$$y_t = A^{-1}\beta_0 + A^{-1}\beta_1 y_{t-1} + A^{-1} u_t \tag{140}$$

or simply:

$$y_t = A_0 + A_1 y_{t-1} + \epsilon_t \tag{141}$$

8.1.4 Block Significance and Causality Tests

When a VAR model includes multiple lags, it can be difficult to determine which variables significantly impact each dependent variable. To address this, **block significance tests** are used to restrict all lags of a particular variable to zero and test their joint significance.

Testing for Block Significance:

Consider a bivariate VAR(3) model:

$$\begin{pmatrix} y_{1t} \\ y_{2t} \end{pmatrix} = \begin{pmatrix} \alpha_{10} \\ \alpha_{20} \end{pmatrix} + \begin{pmatrix} \beta_{11} & \beta_{12} \\ \beta_{21} & \beta_{22} \end{pmatrix} \begin{pmatrix} y_{1t-1} \\ y_{2t-1} \end{pmatrix} + \begin{pmatrix} \gamma_{11} & \gamma_{12} \\ \gamma_{21} & \gamma_{22} \end{pmatrix} \begin{pmatrix} y_{1t-2} \\ y_{2t-2} \end{pmatrix} + \begin{pmatrix} \delta_{11} & \delta_{12} \\ \delta_{21} & \delta_{22} \end{pmatrix} \begin{pmatrix} y_{1t-3} \\ y_{2t-3} \end{pmatrix} + \begin{pmatrix} u_{1t} \\ u_{2t} \end{pmatrix} \tag{142}$$

Expanding into individual equations:

$$\begin{aligned} y_{1t} &= \alpha_{10} + \beta_{11} y_{1t-1} + \beta_{12} y_{2t-1} + \gamma_{11} y_{1t-2} + \gamma_{12} y_{2t-2} \\ &\quad + \delta_{11} y_{1t-3} + \delta_{12} y_{2t-3} + u_{1t}, \end{aligned} \tag{143}$$

$$\begin{aligned} y_{2t} &= \alpha_{20} + \beta_{21} y_{1t-1} + \beta_{22} y_{2t-1} + \gamma_{21} y_{1t-2} + \gamma_{22} y_{2t-2} \\ &\quad + \delta_{21} y_{1t-3} + \delta_{22} y_{2t-3} + u_{2t}. \end{aligned} \tag{144}$$

To test whether a variable affects another, we impose restrictions on its lags. The table below summarizes common hypotheses and their implied restrictions:

Table 2: Granger causality tests and implied restrictions on VAR models

Hypothesis	**Implied Restriction**
Lags of y_{1t} do not explain y_{2t}	$\beta_{21} = 0, \gamma_{21} = 0, \delta_{21} = 0$
Lags of y_{1t} do not explain y_{1t}	$\beta_{11} = 0, \gamma_{11} = 0, \delta_{11} = 0$
Lags of y_{2t} do not explain y_{1t}	$\beta_{12} = 0, \gamma_{12} = 0, \delta_{12} = 0$
Lags of y_{2t} do not explain y_{2t}	$\beta_{22} = 0, \gamma_{22} = 0, \delta_{22} = 0$

Causality Testing and Interpretation:

These hypotheses can be tested using an F-test, comparing the residual sum of squares (RSS) from unrestricted and restricted models.

A related concept is *Granger causality*, which evaluates whether past values of y_1 help predict y_2. The key interpretation rules are:

- If lags of y_1 significantly affect y_2 but not vice versa, y_1 *Granger-causes* y_2, indicating unidirectional causality.
- If lags of both y_1 and y_2 are significant in each other's equations, bi-directional causality (feedback) exists.
- If neither set of lags is significant, y_1 and y_2 are statistically independent.

It is important to note that Granger causality does not imply true causality—it only measures whether past values of one variable improve predictions of another.

Restricted VARs:

A standard VAR model is typically unrestricted, meaning that lags of every variable appear in each equation. While this provides flexibility, it can lead to *highly parameterized models*, consuming degrees of freedom and potentially reducing forecast accuracy.

A *restricted* or *unbalanced* VAR reduces the number of estimated parameters by imposing theoretical or statistical constraints. For example:

- - In a small-large economy model, the small economy's lagged variables might not significantly affect the larger economy, justifying their exclusion.
- Block significance tests (see previous section) may indicate that certain lags do not significantly contribute to the model and can be removed.

However, restrictions should be applied consistently. Removing specific lags of a variable (e.g., excluding y_{2t-1} and y_{2t-3} but keeping y_{2t-2}) is generally not advisable, as a variable should either be included or excluded as a whole.

8.1.5 VARs with Exogenous Variables

Consider the following specification for a VAR(1) where X_t is a vector of exogenous variables and B is a matrix of coefficients

$$y_t = A_0 + A_1 y_{t-1} + BX_t + e_t \tag{7.81}$$

The components of the vector X_t are known as exogenous variables since their values are determined outside of the VAR system – in other words, there are no equations in the VAR with any of the components of X_t as dependent variables. Such a model is sometimes termed a VARX, although it could be viewed as simply a very restricted VAR where there are equations for each of the exogenous variables, but with the coefficients on the RHS in those equations restricted to zero. Such a restriction may be considered desirable if theoretical considerations suggest it, although it is clearly not in the true spirit of VAR modelling, which is not to impose any restrictions on the model but rather to 'let the data decide'.

8.1.6 Impulse Responses and Variance Decompositions

Block F-tests and an examination of causality in a VAR will suggest which of the variables in the model have statistically significant impacts on the future values of each of the variables in the system. But F-test results will not, by construction, be able to explain the sign of the relationship or how long these

effects require to take place. That is, F-test results will not reveal whether changes in the value of a given variable have a positive or negative effect on other variables in the system, or how long it would take for the effect of that variable to work through the system. Such information will, however, be given by an examination of the VAR's impulse responses and variance decompositions.

Impulse responses: trace out the responsiveness of the dependent variables in the VAR to shocks to each of the variables. So, for each variable from each equation separately, a unit shock is applied to the error, and the effects upon the VAR system over time are noted. Effectively, the impulse responses are partial derivatives of the variables $(y_{jt}, j = 1, \ldots, g)$ with respect to each error term $(u_{kt}, k = 1, \ldots, g)$:

$$\frac{\partial y_{jt}}{\partial u_{kt}}. \tag{145}$$

In practice, one standard deviation shocks are often used rather than one unit, as it might be the case that a one-unit shock is empirically implausible, but a one standard deviation shock will almost always be relevant.

If there are g variables in a system, a total of g^2 impulse responses could be generated. The way that this is achieved in practice is by expressing the VAR model as a Vector Moving Average (VMA). That is, consider a VAR(p) process:

$$y_t = A_1 y_{t-1} + A_2 y_{t-2} + \cdots + A_p y_{t-p} + u_t, \tag{146}$$

where u_t is a white noise error term. This can be rewritten in its VMA representation:

$$y_t = \sum_{i=0}^{\infty} \Psi_i u_{t-i}, \tag{147}$$

where the matrices Ψ_i are obtained recursively from the estimated VAR coefficients. The elements of Ψ_i provide the impulse response functions, describing how shocks to the error term u_t propagate over time to affect the endogenous variables y_t.

Provided that the system is stable, meaning that the roots of the characteristic equation lie outside the unit circle, the shock should gradually die away.

To illustrate how impulse responses operate, consider the following bivariate VAR(1)

$$y_t = A_1 y_{t-1} + u_t \tag{148}$$

where

$$A_1 = \begin{bmatrix} 0.5 & 0.3 \\ 0.0 & 0.2 \end{bmatrix} \tag{149}$$

The VAR can also be written out using the elements of the matrices and vectors as

$$\begin{bmatrix} y_{1t} \\ y_{2t} \end{bmatrix} = \begin{bmatrix} 0.5 & 0.3 \\ 0.0 & 0.2 \end{bmatrix} \begin{bmatrix} y_{1t-1} \\ y_{2t-1} \end{bmatrix} + \begin{bmatrix} u_{1t} \\ u_{2t} \end{bmatrix} \tag{150}$$

Consider the effect at time $t = 0, 1, \ldots$, of a unit shock to y_{1t} at time $t = 0$, where the $T \times 1$ vector for y at time t is stacked up and written as simply $[y_t]$ for notational convenience:

$$y_0 = \begin{bmatrix} y_{10} \\ y_{20} \end{bmatrix} = \begin{bmatrix} u_{10} \\ u_{20} \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \tag{151}$$

$$y_1 = A_1 y_0 = \begin{bmatrix} 0.5 & 0.3 \\ 0.0 & 0.2 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 0.5 \\ 0.0 \end{bmatrix} \tag{152}$$

$$y_2 = A_1 y_1 = \begin{bmatrix} 0.5 & 0.3 \\ 0.0 & 0.2 \end{bmatrix} \begin{bmatrix} 0.5 \\ 0.0 \end{bmatrix} = \begin{bmatrix} 0.25 \\ 0.0 \end{bmatrix} \tag{153}$$

and so on. It would thus be possible to plot the impulse response functions of y_{1t} and y_{2t} to a unit shock in y_{1t}. Notice that the effect on y_{2t} is always zero, since the variable y_{1t-1} has a zero coefficient attached to it in the equation for y_{2t}.

Now consider the effect of a unit shock to y_{2t} at time $t = 0$

$$y_0 = \begin{bmatrix} u_{10} \\ u_{20} \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \tag{154}$$

$$y_1 = A_1 y_0 = \begin{bmatrix} 0.5 & 0.3 \\ 0.0 & 0.2 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 0.3 \\ 0.2 \end{bmatrix} \tag{155}$$

$$y_2 = A_1 y_1 = \begin{bmatrix} 0.5 & 0.3 \\ 0.0 & 0.2 \end{bmatrix} \begin{bmatrix} 0.3 \\ 0.2 \end{bmatrix} = \begin{bmatrix} 0.21 \\ 0.04 \end{bmatrix} \tag{156}$$

and so on. Although it is probably fairly easy to see what the effects of shocks to the variables will be in such a simple VAR, the same principles can be applied in the context of VARs containing more equations or more lags, where it is much more difficult to see by eye what are the interactions between the equations.

Variance decompositions offer a slightly different method for examining VAR system dynamics. They give the proportion of the movements in the dependent variables that are due to their 'own' shocks, versus shocks to the other variables. A shock to the ith variable will directly affect that variable of course, but it will also be transmitted to all of the other variables in the system through the dynamic structure of the VAR. Variance decompositions determine how much of the s-step-ahead forecast error variance of a given variable is explained by innovations to each explanatory variable for $s = 1, 2, \ldots$. In practice, it is usually observed that own series shocks explain most of the (forecast) error

variance of the series in a VAR. To some extent, impulse responses and variance decompositions offer very similar information.

For calculating impulse responses and variance decompositions, the ordering of the variables is important. To see why this is the case, recall that the impulse responses refer to a unit shock to the errors of one VAR equation alone. This implies that the error terms of all other equations in the VAR system are held constant. However, this is not realistic since the error terms are likely to be correlated across equations to some extent. Thus, assuming that they are completely independent would lead to a misrepresentation of the system dynamics. In practice, the errors will have a common component that cannot be associated with a single variable alone.

The usual approach to this difficulty is to generate *orthogonalised impulse responses*. In the context of a bivariate VAR, the whole of the common component of the errors is attributed somewhat arbitrarily to the first variable in the VAR. In the general case where there are more than two variables in the VAR, the calculations are more complex but the interpretation is the same. Such a restriction in effect implies an 'ordering' of variables, so that the equation for y_{1t} would be estimated first and then that of y_{2t}, a bit like a recursive or triangular system.

Assuming a particular ordering is necessary to compute the impulse responses and variance decompositions. Ideally, financial theory should suggest an ordering (in other words, that movements in some variables are likely to follow, rather than precede, others). Failing this, the sensitivity of the results to changes in the ordering can be observed by assuming one ordering, and then exactly reversing it and recomputing the impulse responses and variance decompositions. It is also worth noting that the more highly correlated are the residuals from an estimated equation, the more the variable ordering will be important. But when the residuals are almost uncorrelated, the ordering of the variables will make little difference.

8.2 Cointegration (Not in book)

Cointegration describes a relationship between two or more **non-stationary time series** (typically integrated of order 1, or $I(1)$). For a set of series to be cointegrated, two main conditions must be met:

1. They must all be integrated of the **same order** $d > 0$ (e.g., all $I(1)$).
2. There must exist at least one **linear combination** of these series that is **stationary** ($I(0)$).

This stationary linear combination represents a **long-run equilibrium relationship** between the series. Even if the individual series wander over time, they will not drift too far apart from each other.

9 Term Structure Modelling (Not in book)

We first introduce some mathematical concepts often employed for term structure models.

9.1 Principle Components Analysis (PCA)

9.1.1 Introduction

- PCA is a statistical technique used for dimensionality reduction.
- It transforms the original data into a set of uncorrelated variables called *principal components*.
- The first few components capture most of the data's variance, simplifying the dataset while retaining essential information.

9.1.2 Covariance Matrix

Let X be a $n \times p$ data matrix, where n is the number of observations and p is the number of features.

The covariance matrix of X is:

$$\Sigma = \frac{1}{n} X^T X \tag{157}$$

The covariance matrix shows how different features of the data vary with respect to one another.

9.1.3 Eigenvalues and Eigenvectors

PCA is based on the eigenvalue decomposition of the covariance matrix. For covariance matrix Σ, we solve:

$$\Sigma v = \lambda v \tag{158}$$

where:

- v is an eigenvector (principal component).
- λ is the corresponding eigenvalue, indicating the variance captured by that component.

9.1.4 Principle Components

- The principal components are the eigenvectors of the covariance matrix.
- They represent the new coordinate axes along which the data has the most variance.
- The components are ranked by the magnitude of their corresponding eigenvalues.

9.1.5 Dimensionality Reduction

PCA allows us to reduce the number of dimensions while retaining the most significant features.

Keep the top k principal components where k is chosen such that:

$$\frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{p} \lambda_i} \geq 0.95 \tag{159}$$

This preserves 95% of the variance in the data.

Before performing PCA, we center the data by subtracting the mean of each feature. For data matrix X, compute:

$$\tilde{X} = X - \mu \tag{160}$$

where μ is the vector of feature means.

Centering ensures the first principal component corresponds to the direction of maximum variance.

9.1.6 Projection Onto Principle Components

After computing the principal components, we project the original data onto these new axes.

For a data vector x, the projection onto the k-th principal component v_k is:

$$z_k = x^T v_k \tag{161}$$

This results in a lower-dimensional representation of the data.

9.1.7 Interpreting PCA Results

Eigenvalues (λ_i) indicate the amount of variance captured by each principal component.

The proportion of variance explained by the k-th component is:

$$\text{Explained Variance} = \frac{\lambda_k}{\sum_{i=1}^{p} \lambda_i} \tag{162}$$

Cumulative explained variance helps in deciding how many components to retain.

9.2 Singular Value Decomposition (SVD)

9.2.1 Introduction

Singular Value Decomposition (SVD) is a powerful matrix factorization technique used in linear algebra, machine learning, and data science. It decomposes a given matrix A into three matrices that reveal important structural properties of the data.

9.2.2 Mathematical Formulation

Given an $m \times n$ matrix A, SVD decomposes it into three matrices:

$$A = U\Sigma V^T \tag{163}$$

where:

- U is an $m \times m$ orthogonal matrix whose columns are the **left singular vectors** of A.
- Σ is an $m \times n$ diagonal matrix containing **singular values**, which represent the strength of each principal component.
- V^T is the transpose of an $n \times n$ orthogonal matrix V, whose columns are the **right singular vectors** of A.

9.2.3 Interpretation and Applications

- **Dimensionality Reduction**: By keeping only the largest singular values, we can approximate A with a lower-rank matrix.
- **Principal Component Analysis (PCA)**: The right singular vectors in V correspond to principal components.

9.3 Comparison: PCA vs. SVD in Financial Applications

- **PCA is typically computed via SVD**: In financial applications, PCA is often performed using SVD for better numerical stability.
- **PCA focuses on variance**, while SVD provides a broader decomposition that is useful for data with missing values.
- **PCA assumes data is centered**, while SVD can be applied to raw financial datasets without centering.
- **PCA is mainly used for risk modeling**, whereas SVD is useful for broader applications like market microstructure analysis.

9.4 Definition of Term Structures

In finance, a **term structure** describes the relationship between the yield (or price, or volatility) of a financial instrument and its **time to maturity**, holding all other characteristics of the instrument (such as credit risk) constant. It essentially provides a snapshot of how a particular financial variable is priced or valued across different future time horizons.

9.4.1 Examples

Term structures are prevalent across various financial markets:

- **Interest Rate Term Structure (Yield Curve):** This is perhaps the most widely known term structure. It plots the yields of bonds (or interest rate swaps) with similar credit quality against their respective maturities (e.g., 3-month, 1-year, 5-year, 10-year, 30-year). The graphical representation is commonly referred to as the yield curve.
- **Commodity Futures Term Structure:** This shows the prices of futures contracts for a specific commodity (e.g., crude oil, gold, agricultural products) for different delivery dates in the future. The shape of this curve (e.g., contango or backwardation) reflects market expectations of future supply and demand, storage costs, and convenience yields.
- **Volatility Term Structure:** This illustrates the implied volatility derived from options prices on a particular underlying asset (e.g., a stock index, a currency pair) for different option expiration dates. It reflects the market's expectation of future price fluctuations of the underlying asset over various time horizons.

9.5 The Importance of Term Structure Modelling

Understanding and modelling term structures is crucial in economics and finance for several reasons:

9.5.1 Economic Interpretations

The shape of a term structure, particularly the yield curve, offers valuable economic insights:

- **Level:** The overall height of the curve (average yield) often reflects the current stance of monetary policy, prevailing inflation rates, and general expectations about future inflation.
- **Slope:** The difference between long-term and short-term rates. A positively sloped (upward sloping) curve, where long-term rates are higher than short-term rates, is typical and often suggests expectations of economic expansion and/or rising inflation. Conversely, an inverted (downward sloping) curve can be a leading indicator of a potential economic recession.

- **Curvature:** The "bow" or "hump" in the curve. This can reflect market segmentation, specific supply/demand pressures at certain maturities, or more complex expectations about the path of future interest rates.

9.5.2 Predictive Insights and Risk Premia

The economic interpretations highlight that term structures carry information about the future economic environment. For instance, the yield on a long-term bond can be conceptually decomposed into two main parts:

1. The average of expected future short-term interest rates over the life of the bond.
2. A **term premium** (or risk premium), which is the additional compensation investors demand for bearing the risks associated with holding a longer-maturity instrument (such as interest rate risk or inflation uncertainty).

Modelling these components helps in understanding market expectations and risk appetites.

9.5.3 Diverse Applications

Term structure models are utilized by a wide range of participants:

- **Macro-economists** use them to gauge current economic conditions, assess monetary policy effectiveness, and forecast economic activity.
- **Financial economists** develop theoretical models to explain the behavior and dynamics of term structures.
- **Market participants** (e.g., traders, portfolio managers, risk managers) employ these models for pricing interest rate derivatives, hedging interest rate risk, constructing investment strategies, and valuing fixed-income securities.

The varying needs of these users have led to the development of a diverse array of term structure models, differing in their assumptions, mathematical form, and empirical fit.

9.6 Factor Models: A General Overview

9.6.1 Core Concept and Purpose

A **factor model** is a statistical model that attempts to explain the behavior of a potentially large number of observed variables in terms of a smaller number of unobserved (or sometimes observed) underlying variables, known as **factors**. The primary purposes of factor models include:

- **Dimensionality Reduction:** Simplifying complex systems by representing the co-movement of many variables with fewer factors.
- **Explanation:** Identifying the key common drivers of variation within a set of variables.
- **Prediction:** Forecasting the behavior of the observed variables based on forecasts of the underlying factors.
- **Risk Management:** Understanding how exposure to different systematic factors contributes to overall portfolio or asset risk.

9.6.2 Types of Factors

Factors can be broadly categorized as:

- **Observable Factors:** These are directly measurable economic or financial variables, such as GDP growth, inflation rates, market indices, or specific commodity prices.
- **Latent (Unobserved) Factors:** These are statistical constructs derived from the observed data itself (e.g., through techniques like Principal Component Analysis). They represent common patterns of co-movement but may not have an immediate, pre-defined economic label.

9.7 Term Structure Factor Models

9.7.1 Applying Factor Models to Term Structures

In the context of term structures, the "high-dimensional object" is the set of yields (or prices/volatilities) across many different maturities. A term structure factor model aims to explain the movements of this entire curve using a small number of common underlying factors. This is based on the empirical observation that yields at different maturities tend to move together, suggesting the presence of common drivers. Many influential term structure models assume that the dynamics of interest rates are driven by a few latent factors. These factors can be entirely unobserved (statistically filtered from observed yields), based on observable variables (like specific benchmark yields or macroeconomic variables), or a combination thereof.

9.7.2 Dynamic Term Structure Models (DTSMs) - Conceptual Components

Affine Dynamic Term Structure Models are a prominent class of factor models. Conceptually, a discrete-time affine DTSM typically involves three key elements:

1. **Risk Factors (X_t):** A vector of N (where N is typically a small number, e.g., 1 to 3) underlying risk factors that are assumed to drive the entire term structure. These factors are often modelled as following a

first-order Vector Autoregressive (VAR) process under the objective (real-world) probability measure P. This means their future values depend linearly on their current values, plus some random shocks.

2. **Short-Term Interest Rate (r_t):** The instantaneous or very short-term risk-free interest rate is assumed to be an affine (linear plus a constant) function of these pricing factors. This provides a direct link between the unobserved factors and an observable point on the yield curve.

3. **Stochastic Discount Factor (SDF, M_{t+1}):** A theoretical construct used for asset pricing. The SDF prices all assets under the absence of arbitrage. In affine DTSMs, the logarithm of the SDF is typically specified as an affine function of the state factors and the short-term rate, incorporating the market price of risk associated with the factors.

The Affine Property The "affine" property in these models means that bond yields (log bond prices divided by maturity) are also affine functions of the state variables (factors). This mathematical structure is highly convenient as it often leads to closed-form or easily computable solutions for bond prices and yields, making the models tractable for estimation and application.

9.8 Principal Component Analysis (PCA) in Term Structure Modelling

9.8.1 What is PCA?

Principal Component Analysis (PCA) is a statistical technique used for dimensionality reduction. It transforms a dataset of possibly correlated variables into a new set of linearly uncorrelated variables called **principal components (PCs)**. These PCs are ordered such that:

- The first principal component (PC1) accounts for the largest possible variance in the original data.
- The second principal component (PC2) accounts for the largest possible amount of the remaining variance, subject to being uncorrelated with PC1.
- This continues for subsequent components, each capturing successively smaller portions of the remaining variance and being uncorrelated with all preceding components.

9.8.2 PCA as a Factor Identification Tool

In term structure modelling, PCA is widely used to identify and construct the latent factors that drive the movements of the yield curve (or other term structures). The first few principal components of the observed yield changes (or levels) are often used as estimates of the underlying unobserved factors (X_t) in factor models. For a given data sample of yields y_t, the first p principal components can be written as a linear combination $X_t = Ly_t$, where L is a matrix of loadings.

9.8.3 Economic Interpretation of Principal Components

A significant finding, notably highlighted by Litterman and Scheinkman (1991) for U.S. Treasury yields, is that the first few principal components of yield curve changes often have intuitive economic interpretations:

1. **Level Factor (PC1):** This component typically represents parallel shifts in the entire yield curve. A change in PC1 implies that all yields, from short-term to long-term, tend to move up or down together by roughly the same amount. The loadings associated with PC1 are usually fairly constant across maturities.

2. **Slope Factor (PC2):** This component captures changes in the steepness of the yield curve. A change in PC2 affects short-term and long-term yields differently (often in opposite directions or with significantly different magnitudes), thus altering the slope. The loadings for PC2 typically have opposite signs or vastly different magnitudes at the short and long ends of the maturity spectrum.

3. **Curvature Factor (PC3):** This component represents changes in the concavity or "bow" of the yield curve. A change in PC3 might cause medium-term yields to move differently relative to short-term and long-term yields. The loadings for PC3 often exhibit a "humped" or "U" shape, affecting mid-maturities differently from the ends.

9.8.4 Empirical Significance and Sufficiency of Factors

Empirically, these first three principal components are remarkably effective at explaining the variation in yield curve movements.

- PC1 (Level) often accounts for the vast majority (e.g., 80-95%) of the total variance.

- PC2 (Slope) typically explains a significant portion of the remaining variance (e.g., 5-15%).

- PC3 (Curvature) accounts for a smaller, but often still relevant, portion of the variance (e.g., 1-5%).

Together, these three factors frequently explain over 95% (and often over 99%) of the total variation in yield curve changes. This empirical regularity suggests that three factors are often sufficient to capture most term structure dynamics. However, for specific applications or markets (e.g., commodity futures with strong seasonality, or markets with complex autocorrelation patterns), more factors might be necessary. Similar PCA-based factor interpretations (level, slope, curvature) are often found for other term structures, such as commodity futures price curves or implied volatility term structures, although the relative importance of the factors can vary.

9.9 Broader Classes of Term Structure Models

While factor models, particularly affine DTSMs using PCA-derived factors, are a cornerstone of term structure analysis, it's useful to be aware of broader model classifications:

1. **Pricing (No-Arbitrage) Models (e.g., Heath-Jarrow-Morton (HJM), Hull-White):**
    - **Focus:** Perfectly fitting the observed term structure at a specific point in time to ensure that no arbitrage opportunities exist.
    - **Primary Use:** Pricing and hedging of interest rate derivatives.
    - **Sub-classes:**
        - *Fundamental models:* Involve joint specification of risk premia and real-world dynamics.
        - *Reduced-form models:* Directly specify the evolution of the forward rate curve or yield curve under a risk-neutral probability measure (used for pricing).
2. **Structural (Equilibrium) Models (e.g., Vasicek, Cox-Ingersoll-Ross (CIR), Affine Class (ACM, JSZ)):**
    - **Focus:** Modelling the dynamics of the term structure based on assumptions about the underlying economy, the behavior of economic agents, and the determination of risk premia.
    - **Characteristics:** Often aim to derive the term structure from more fundamental economic variables or stochastic processes. Affine factor models (DTSMs) typically fall into this category or bridge between structural and pricing approaches.
3. **Cointegrated Models:**
    - **Focus:** Model the term structure as a system where multiple yields (which are individually non-stationary) are cointegrated. This implies that there are long-run equilibrium relationships between yields at different maturities.
    - **Characteristics:** Typically involve one underlying common stochastic trend, with spreads between different yields being stationary around this trend.

Factor models often share characteristics with, or can be classified within, both pricing and structural/equilibrium approaches, depending on their specific formulation and intended application.

10 Panel Data Analysis

10.1 What is Panel Data?

- Combines cross-section and time-series data.
- The same cross-sectional unit (industry, firm, country) is surveyed over time.
- Examples:
  - Time series: Data observed at different points in time.
  - Cross-section: Data observed at a single point in time.

$$X = \begin{bmatrix} x_{1,1} & x_{1,2} & x_{1,3} & \dots & x_{1,T} \\ x_{2,1} & x_{2,2} & x_{2,3} & \dots & x_{2,T} \\ x_{3,1} & x_{3,2} & x_{3,3} & \dots & x_{3,T} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{N,1} & x_{N,2} & x_{N,3} & \dots & x_{N,T} \end{bmatrix}$$

a matrix of panel data observations on variable X, N cross-sectional observations, T time series observations.

10.2 Why Use Panel Data?

- We can address a broader range of issues and tackle more complex problems with panel data.
- It is often of interest to examine how variables, or the relationships between them, change dynamically (over time).
  - To do this using pure time series data would often require a long run of data simply to get a sufficient number of observations to be able to conduct any meaningful hypothesis tests.
  - The additional variation can also help to mitigate problems of multicollinearity.
- By structuring the model in an appropriate way, we can remove the impact of certain forms of omitted variable bias in regression results.

10.2.1 Different Types of Variation

- **Regressors**
  - Varying regressors (annual income for a person, annual consumption of a product)
  - Time-invariant regressors (country, gender, race, education)

  - Individual-invariant regressors (time trend, economy trends such as unemployment rate, market index)

- **Overall variation**: variation over time and individuals.
  - **Between variation**: variation between individuals.
  - **Within variation**: variation within individuals (over time).

10.2.2 Panel Data Decomposition Table

Id i	Time t	Variable x_{it}	Ind. Mean $\bar{x}_i$	Overall Mean $\bar{x}$	Overall Dev. $x_{it} - \bar{x}$	Between Dev. $\bar{x}_i - \bar{x}$	Within Dev. $x_{it} - \bar{x}_i$
1	1	9	10	20	-11	-10	-1
1	2	10	10	20	-10	-10	0
1	3	11	10	20	-9	-10	1
2	1	20	20	20	0	0	0
2	2	20	20	20	0	0	0
2	3	20	20	20	0	0	0
3	1	25	30	20	5	10	-5
3	2	30	30	20	10	10	0
3	3	35	30	20	15	10	5

10.3 Econometric Models

Econometrically, the simplest setup is

$$y_{it} = \alpha + \beta x_{it} + u_{it}$$

where y_{it} is the dependent variable, α is the intercept term, β is a slope parameter to be estimated on the explanatory variable, x_{it}; $t = 1, \ldots, T$; $i = 1, \ldots, N$.

10.3.1 Pooled Regression

The simplest way to deal with this data would be to estimate a single, pooled regression on all the observations together.

Person	Year	Income	Age	Sex
1	2001	1300	27	1
1	2002	1600	28	1
1	2003	2000	29	1
2	2001	2000	38	2
2	2002	2300	39	2
2	2003	2400	40	2

But pooling the data assumes that there is no heterogeneity – i.e., the same relationship holds for all the data.

- The constant, α, is the same across groups and time.
- The coefficient, β, is constant across groups and time.
- Any differences across groups enter the model only through the error term.

But, pooled regression may result in heterogeneity bias:

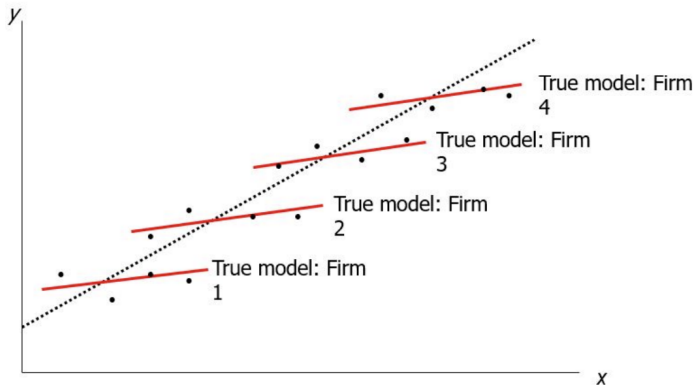


Figure 4: The figure illustrates *heterogeneity bias* in pooled regression. The true relationship between x and y differs across firms, as indicated by the separate red lines for each firm. However, the pooled regression (dotted black line) assumes a single relationship across all firms, ignoring firm-specific differences. This leads to a biased estimate of the true relationship, as the slope and intercept are incorrectly averaged across heterogeneous groups.

10.4 Key Concept: Unobserved Heterogeneity

In panel data models, we assume that there is unobserved heterogeneity across individuals. For example, a person´s unobserved ability (e.g., intelligence, soft skills and motivation) affects wages but is not explicitly included in the dataset. The main question that distinguishes Fixed Effects and Random Effects model is how they treat individual-specific effects.

10.5 Fixed vs Random Effects Model

10.5.1 Preview:

Fixed Effects (FE) Model:

- Assumes individual-specific characteristics may be **correlated** with the explanatory variables.

- The heterogeneity is captured by an **intercept** that varies for each individual.
- The model removes bias by **differencing out** or **demeaning** the individual-specific effects.
- **Downside**: Does not allow for time-invariant variables (e.g., gender, education) as regressors.

Random Effects (RE) Model:

- Assumes individual-specific characteristics are **NOT correlated** with the explanatory variables.
- The heterogeneity enters the **error term**, treated as a random variable.
- More efficient estimation since individual-specific means do not need to be removed.
- **Downside**: If the assumption of no correlation is incorrect, estimates will be biased.

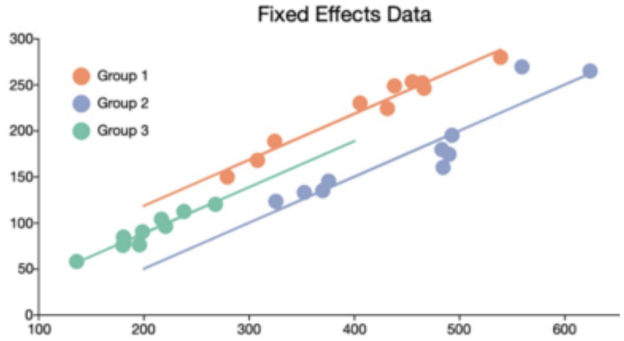
10.5.2 Fixed Effects Model

- In **fixed effects models**, we assume groups share **common coefficients** on regressors but have **group-specific intercepts**.
- Model:

$$y_{it} = \alpha + \beta X_{it} + \mu_i + \nu_{it}$$

- μ_i accounts for unobserved heterogeneity across individuals.

We can visualize fixed effects data as follows:



Fixed Effects Model Assumptions

1. $y_{it} = \alpha + X_{it}\beta + \mu_i + \nu_{it}, \quad i = 1, \ldots, N, \quad t = 1, \ldots, T$
2. μ_i is an individual-specific (fixed) effect, possibly correlated with X_{it}.
3. $E[\nu_{it}] = 0, \quad \forall i, t$
4. $E[\nu_{it}^2] = \sigma_\nu^2$ (homoscedastic idiosyncratic error)
5. $E[\nu_{it}\nu_{js}] = 0 \quad (i \neq j)$ or $(t \neq s)$ (no autocorrelation)
6. $E[\nu_{it} \mid X_{is}] = 0, \quad \forall i, t, s$ (exogeneity of ν_{it})

Individual-Specific (fixed) Effect:

- Example: The sector a firm operates in, a person's gender, or the country where a bank has its headquarters.
- μ_i encapsulates all variables that affect y_{it} cross-sectionally but do not vary over time.
- Individual-specific effects can be recovered as $\bar{y}_i - \alpha - \beta\bar{x}_i$.
- The fixed effects model can be estimated in two ways:
  - using dummy variables, known as the **Least Squares Dummy Variable (LSDV) approach**, or
  - using **Fixed Effects Within Estimator**.

Within vs Between Estimation:
Within Transformation:

- The within transformation involves subtracting the time-mean of each entity away from the values of the variable.

$$y_{it} - \bar{y}_i = \beta(X_{it} - \bar{X}_i) + \nu_{it} - \bar{\nu}_i$$

- Note that such a regression does not require an intercept term since now the dependent variable will have zero mean by construction.
- We can write it as:

$$\ddot{y}_{it} = \beta\ddot{X}_{it} + \ddot{\nu}_{it}$$

- where the double dots above the variables denote the demeaned values.
- This model can be estimated using OLS, but we need to make a degrees of freedom correction.
  - Because you remove one additional parameter per cross-sectional unit (namely, its mean).

Between Estimation:

- Instead of demeaning the data (within estimator), an alternative is to run a **cross-sectional regression** on the **time-averaged values** of the variables. This is called the **Between Estimator**.

$$\bar{y}_i = \alpha + \beta \bar{x}_i + \mu_i + \bar{\nu}_i \tag{9}$$

- Regression is run on the average values of each variable.
- Used when focusing on **cross-sectional variation**.
- Helps mitigate measurement error by averaging over time.
- A between estimator does not omit time-invariant variables like a within estimator, but you have to assume no unobserved heterogeneity between individuals.
- In this case, a random effects estimator which combines both within and between group variation is usually more efficient.

Key insights:

- The Between Estimator uses differences across individuals but does not remove fixed individual effects
- The Fixed Effects Model (Within Estimator) relies on within-individual variation by removing the mean

Clustering Standard Erros:

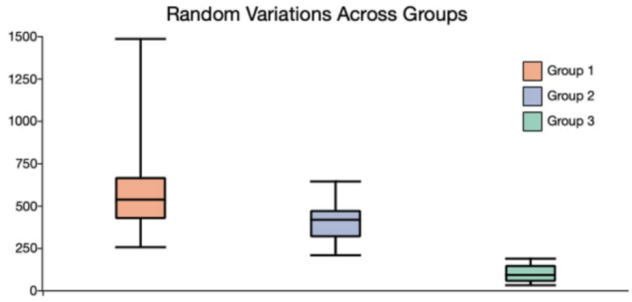
- **Within-Entity Correlation:** In a fixed effects model, each entity (individual, firm, country, etc.) tends to have error terms that are correlated over time—even after demeaning or differencing—because all time-invariant effects are absorbed into the fixed effect but any remaining time-specific shocks for the same entity can still be correlated.
- **Why Cluster?** Clustering the standard errors at the entity level corrects for this serial correlation within each entity.
- Without clustering (or another robust correction), standard errors are often underestimated, leading to:
  - Inflated test statistics.
  - Overly narrow confidence intervals.
  - Increased Type I error risk.
- **Takeaway:** For valid inference in fixed effects models, it is standard practice to cluster standard errors at the entity level.

10.5.3 Random Effects Model

- An alternative to the fixed effects model described above is the random effects model, which is sometimes also known as the error components model.
- Again, assumes different intercept terms for each entity that are constant over time.
- **BUT!** The difference is that the intercepts for each cross-sectional unit are assumed to arise from a common intercept α plus a random variable ε_i that varies cross-sectionally but is constant over time.
- Model:

$$y_{it} = \alpha + \beta X_{it} + \omega_{it}, \quad \omega_{it} = \varepsilon_i + \nu_{it} \tag{14}$$

- Requires strict assumptions about error term independence.
- ε_i measures the random deviation of each entity's intercept term from the **global** intercept, while ν_{it} represents the idiosyncratic error term.
- In Random Effects Model ε_i measures the random deviation of each entity´s intercept term from the global intercept term α



- Within-group variation is shown by each boxplot's spread.
- Between-group variation is evident from the differences in medians and overall ranges across the groups.
- A random effects model would capture these group-to-group differences by treating each group's effect as a random draw from a common distribution, where each group has its own *level* and variability.

Model Assumptions:

- Unlike the fixed effects model, there are no dummy variables to capture the heterogeneity (variation) in the cross-sectional dimension.

- Instead, this occurs via the error terms, which requires strict assumptions.

$$y_{it} = \alpha + X_{it}\beta + \varepsilon_i + \nu_{it}, \quad i = 1, \ldots, N, \quad t = 1, \ldots, T$$

1. $E[\varepsilon_i] = E[\nu_{it}] = 0, \quad \forall i, t$
2. $E[\varepsilon_i^2] = \sigma_\varepsilon^2, \quad E[\nu_{it}] = \sigma_\nu^2, \quad$ (both components homoscedastic)
3. $E[\varepsilon_i \nu_{jt}] = 0, \quad \forall i, t \quad$ (independence of two components)
4. $E[\varepsilon_i \varepsilon_j] = 0 \quad$ if $i \neq j \quad$ (no across-group correlation)
5. $E[\nu_{it}\nu_{js}] = 0 \quad$ if $t \neq s$ or $i \neq j \quad$ (no autocorrelation)
6. $E[\varepsilon_i \mid X_{is}] = E[\nu_{it} \mid X_{is}] = 0, \quad \forall i, t, s \quad$ (independent of regressors)

Model Estimation

- **Model Setup:**

  We consider the following panel-data model:

$$y_{it} \;=\; \alpha \;+\; X_{it}\,\beta \;+\; \omega_{it}, \quad \omega_{it} \;=\; \varepsilon_i \;+\; \nu_{it},$$

  where:

  - ε_i is the *individual (random) effect*, common to all time periods t for unit i.
  - ν_{it} is the *idiosyncratic* (time-specific) error term for unit i.

- **Covariance Decomposition:**

  Because ε_i is constant over time for each i, the composite error ω_{it} is correlated across different time periods t and s (with $t \neq s$). In particular,

$$\mathrm{Cov}\big(\omega_{it}, \omega_{is}\big) \;=\; \mathrm{Cov}\big(\varepsilon_i + \nu_{it},\, \varepsilon_i + \nu_{is}\big) \;=\; \mathbb{E}\big[\varepsilon_i^2\big] \;=\; \sigma_\varepsilon^2 \;>\; 0.$$

  This nonzero correlation (for $t \neq s$) violates the usual OLS assumption that errors are uncorrelated across observations.

- OLS estimation of (α, β) remains *consistent*, but OLS is no longer *efficient* because it ignores the correlation structure in ω_{it}. Moreover, the usual OLS standard errors must be adjusted for within-unit serial correlation.

- To handle this correlation properly, we can use **Generalized Least Squares (GLS)**. In practice, one often applies a *quasi-demeaning* (or "random effects") transformation to simplify the correlation structure before estimating by GLS.

GLS Transformation (Quasi-Demeaning)

- **Transformation:**

  Define the transformed variables:

  $$\tilde{y}_{it} \;=\; y_{it} - \theta\,\overline{y}_i, \quad \tilde{X}_{it} \;=\; X_{it} - \theta\,\overline{X}_i, \quad \text{where} \quad \overline{y}_i \;=\; \frac{1}{T}\sum_{t=1}^{T} y_{it}, \;\; \overline{X}_i \;=\; \frac{1}{T}\sum_{t=1}^{T} X_{it}.$$

  The constant $0 \le \theta \le 1$ is chosen according to

  $$\theta \;=\; 1 \;-\; \sqrt{\frac{\sigma_\nu^2}{\sigma_\nu^2 \;+\; T\,\sigma_\varepsilon^2}}.$$

  Here $\sigma_\nu^2 = \mathrm{Var}(\nu_{it})$ and $\sigma_\varepsilon^2 = \mathrm{Var}(\varepsilon_i)$.

- **Random Effects (RE) Estimator:**

  After this transformation, one estimates

  $$\tilde{y}_{it} \;=\; \alpha(1-\theta) \;+\; \tilde{X}_{it}\,\beta \;+\; \tilde{\omega}_{it} \quad \text{via OLS},$$

  where

  $$\tilde{\omega}_{it} \;=\; \big(\varepsilon_i + \nu_{it}\big) \;-\; \theta\big(\varepsilon_i + \overline{\nu}_i\big) \;=\; (1-\theta)\,\varepsilon_i + \nu_{it} - \theta\,\overline{\nu}_i,$$

  and $\overline{\nu}_i = \frac{1}{T}\sum_{t=1}^{T} \nu_{it}$. The resulting "RE" estimator can be viewed as a weighted average of:

  - *Between* (or OLS on group means) estimation, and
  - *Within* (fixed-effects) estimation.

  Specifically,

  - If $\theta = 1$, we fully difference out ε_i, reproducing the Fixed Effects estimator.
  - If $\theta = 0$, we use OLS on the untransformed data.
  - For $0 < \theta < 1$, we partially remove ε_i, retaining some between-group variation.

Implementation Notes:

- **Effect of the Quasi-Demeaning Transformation:** The quasi-demeaning transformation *reduces*, but does not fully eliminate, cross-correlations in the error terms. The transformed error term still exhibits some correlation across time due to the presence of the individual random effect. Specifically, for $t \neq s$, we have:

  $$\mathbb{E}[\tilde{\omega}_{it}\tilde{\omega}_{is}] = (1-\theta)^2\sigma_\varepsilon^2 + \theta^2\frac{\sigma_\nu^2}{T}.$$

  Thus, while the transformation simplifies the correlation structure, it does not completely remove it.

- **Efficiency of the RE Estimator:** The Random Effects (RE) estimator is *fully efficient* under the assumption that the individual effect ε_i is uncorrelated with the regressors X_{it}. Under this assumption, the Generalized Least Squares (GLS) estimator obtained after quasi-demeaning is the Best Linear Unbiased Estimator (BLUE), achieving higher efficiency than Ordinary Least Squares (OLS). However, if ε_i is correlated with X_{it}, then the RE estimator becomes inconsistent, and the Fixed Effects (FE) estimator would be preferred.

10.5.4 Fixed vs Random Effects Comments

- The Random Effects (RE) model is often considered more appropriate when the entities in the sample can be thought of as having been randomly selected from a larger population. In contrast, the Fixed Effects (FE) model is more plausible when the entities in the sample effectively constitute the entire population of interest.
- More technically, the transformation involved in the Generalized Least Squares (GLS) procedure under the RE approach does not remove explanatory variables that do not vary over time. This is a key advantage because it allows for the estimation of time-invariant effects, which are differenced out in the FE model.
- Since the RE model does not require the inclusion of dummy variables or a within transformation, fewer parameters need to be estimated. This saves degrees of freedom and makes the RE estimator more efficient than the FE estimator when its assumptions hold.
- However, a major drawback of the RE approach is that it relies on the assumption that the composite error term is uncorrelated with all explanatory variables. If this assumption is violated, the RE estimator becomes inconsistent, and the FE model should be preferred.

10.5.5 Models and Estimators

Estimator / True Model	**Pooled Model**	**RE Model**	**FE Model**
Pooled OLS estimator	Consistent	Inconsistent	Inconsistent
Between estimator	Consistent	Consistent	Inconsistent
Within (FE) estimator	Consistent	Consistent	Consistent
Random effects estimator	Consistent	Consistent	Inconsistent

Table 3: Consistency of Different Estimators Under Different True Models

- The **Fixed Effects (FE) estimator** is always *consistent* under the classical assumptions, whether the true model is pooled, random effects, or fixed effects. However, if the RE assumptions hold, FE is less *efficient* than the RE estimator.

- The **Random Effects (RE) estimator** is *consistent* if the individual effects are uncorrelated with the regressors (the RE assumption), which includes the pooled model as a special case (when the variance of the individual effects is zero). If the individual-specific effects are correlated with the regressors (i.e., a true FE scenario), then the RE estimator is *inconsistent*.
- The **Pooled OLS estimator** is *consistent* only when there is no unobserved heterogeneity (i.e., the pooled model). If the true model is either RE (nonzero random effects) or FE (correlated effects), Pooled OLS does not account for the individual-specific component and is *inconsistent*.
- The **Between estimator** is *consistent* under the pooled and RE assumptions but is *inconsistent* if the individual-specific effects are correlated with regressors (the FE case).
- When the RE assumptions are valid, the RE estimator is more *efficient* than FE; however, if those assumptions are violated, FE remains consistent and is therefore preferred despite potentially lower efficiency.

10.5.6 Hausman Test

Key idea:

- The fixed effects (FE) estimator is more accurate (unbiased under correlation) but less efficient.
- The random effects (RE) estimator is more efficient (smaller variance) but may be biased if ε_i is correlated with regressors.

Hausman specification test:

- Compares parameter estimates under FE ($\hat{\beta}_{FE}$) and RE ($\hat{\beta}_{RE}$).
- Null hypothesis: $H_0 : \hat{\beta}_{FE} = \hat{\beta}_{RE}$.
- If we reject H_0, then the RE estimator is inconsistent, so FE is preferred.
- If we fail to reject H_0, then RE is consistent (and more efficient).

Test statistic:

$$H = (\hat{\beta}_{FE} - \hat{\beta}_{RE})' \left[\text{Var}(\hat{\beta}_{FE}) - \text{Var}(\hat{\beta}_{RE}) \right]^{-1} (\hat{\beta}_{FE} - \hat{\beta}_{RE}),$$

which follows a χ^2 distribution with k degrees of freedom, where k is the number of estimated parameters.

10.5.7 Goodness of Fit

Define variance decompositions:

$$\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}(y_{it}-\bar{y})^2 = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}(y_{it}-\bar{y}_i)^2 + \frac{1}{N}\sum_{i=1}^{N}(\bar{y}_i-\bar{y})^2$$

$$\text{total variance} = \text{within variance} + \text{between variance}$$

Measures of R^2:

$$R^2_{\text{overall}}(\hat{\beta}_{RE}) = \Big[\text{corr}(x_{it}\hat{\beta}_{RE}, y_{it})\Big]^2$$

$$R^2_{\text{between}}(\hat{\beta}_{BE}) = \Big[\text{corr}(\bar{x}_i\hat{\beta}_{BE}, \bar{y}_i)\Big]^2$$

$$R^2_{\text{within}}(\hat{\beta}_{FE}) = \Big[\text{corr}\Big((x_{it}-\bar{x}_i)\hat{\beta}_{FE}, y_{it}-\bar{y}_i\Big)\Big]^2$$

11 Financial Risk Management and Portfolio Optimization

11.1 Introduction

11.1.1 Risk

- All decisions undertaken by financial institutions that may lead to uncertain results can be seen as an element of risk;
- In finance, we generally talk about risk and uncertainty indiscriminately;

11.1.2 Risk Measurement

Risk measurement is a crucial component of financial risk management, involving the quantification of risk in monetary terms to ensure sufficient capital buffers.

11.1.3 Introductory Concepts

Notation:

1. X - Portfolio's value at time one
2. R_0 - Percentage return of risk-free asset
3. ρ - Risk measure

Definition:
A risk measure, ρ, is a function that transforms the random variable X into a single number, $\rho(X)$, that represents the risk associated with the portfolio.

Interpretation:
If at time zero, we model the value of the portfolio at time one with X, then $\rho(X)$ can be seen as the minimum capital required that needs to be added to the initial position (invested in the risk-free instrument) to make the final position acceptable.

- $\rho(X) \leq 0$ then the portfolio is acceptable; no capital needs to be added to the initial position;
- $\rho(X) > 0$ then at least the amount $\rho(X)$ has to be invested in the risk-free instrument.

11.2 Desirable Properties for Risk Measures

11.2.1 Monetary Risk Measures

1. Monotonicity: If $X_2 \leq X_1$ then $\rho(X_2) \geq \rho(X_1)$

- If investment 1 surely has a future value larger than investment 2, then investment 2 must be riskier than investment 1.
- In order to make the two investments acceptable, the cash amount that has to be added to the second position is larger than the one needed for the first position.

2. Cash Invariance: $\rho(X + cR_0) = \rho(X) - c$

- Adding cash reduces the risk of the initial position in the same amount.
- Also called translational invariance.

Definition:
A risk measure satisfying the monotonicity and cash invariance properties is called a monetary risk measure.

11.2.2 Coherent Risk Measures

3. Positive homogeneity: $\rho(\lambda X) = \lambda\rho(X)$

- Risk increases linearly with the size of the position;
- If we double the position size, then the risk doubles as well.

4. Subadditivity: $\rho(X_1 + X_2) \leq \rho(X_1) + \rho(X_2)$

- Diversification should be rewarded;

- Joining different positions decreases the total risk. The sum of the individual risks bounds the risk of the joint position;
- Regulators requiring the use of subadditive risk measures discourage financial institutions from splitting into different parts to have less capital requirements.

Definition:
A risk measure satisfying the monotonicity, cash invariance, positive homogeneity, and subadditivity is called a coherent risk measure.

Issues with Coherent Risk Measures:

- Positive homogeneity means that risk increases linearly with the size of the position, which is not always true
- it ignores the liquidity risk.

11.2.3 Convex Risk Measures

5. Convexity: $\rho(\lambda X_1 + (1-\lambda)X_2) \leq \lambda\rho(X_1) + (1-\lambda)\rho(X_2)$, for $\lambda \in [0,1]$

- Investment 1 leads to a value X_1 and investment 2 leads to a value X_2. Diversifying and investing in both a proportion of our capital leads to a value $\lambda X_1 + (1-\lambda)X_2$;
- A risk measure should reward diversification!
- Note that if ρ is a convex risk measure then

$$\rho(\lambda X_1 + (1-\lambda)X_2) \leq \max(\rho(X_1), \rho(X_2))$$

Definition:
A convex risk measure is a risk measure satisfying the monotonicity, cash invariance, and convexity properties.

11.2.4 Coherent vs Convex Risk Measures

Equivalent Definition of Coherent Risk Measure:
A convex risk measure that satisfies the positive homogeneity property is called a coherent risk measure.

- Convexity together with positive homogeneity implies subadditivity.

$$\begin{aligned}\rho(X_1 + X_2) &= \rho\left(\lambda\frac{1}{\lambda}X_1 + (1-\lambda)\frac{1}{1-\lambda}X_2\right)\\ &\leq \lambda\rho\left(\frac{1}{\lambda}X_1\right) + (1-\lambda)\rho\left(\frac{1}{1-\lambda}X_2\right)\\ &\leq \rho(X_1) + \rho(X_2)\end{aligned}$$

- Subadditivity together with positive homogeneity implies convexity.
- Convexity alone does not imply positive homogeneity and subadditivity.

11.2.5 Mean-Variance Risk

We define the *mean–variance risk measure* ρ by

$$\rho(X) \;=\; -\,\mathbb{E}\Big(\tfrac{X}{R_0}\Big) \;+\; c\sqrt{\mathrm{Var}\Big(\tfrac{X}{R_0}\Big)},$$

where $R_0 > 0$ is a constant (often taken to be 1 for simplicity) and $c > 0$ is a fixed parameter controlling the relative weight of the variance term.

Properties Satisfied:
The measure satisfies Cash invariance, Positive homogeneity and Sub-additivity. However, it does not satisfy monotonicity, and thus is neither coherent or convex

11.2.6 Variance vs Standard Deviation

Standard deviation is subadditive, whilst variance is not. We can see this from the relations:

$$\sqrt{\mathrm{Var}(X+Y)} \;\le\; \sqrt{\mathrm{Var}(X)} \;+\; \sqrt{\mathrm{Var}(Y)}.$$

$$\mathrm{Var}(X+Y) \;=\; \mathrm{Var}(X) \;+\; \mathrm{Var}(Y) \;+\; 2\,\mathrm{Cov}(X,Y).$$

11.3 Value-at-risk (VaR)

Value-at-Risk (VaR) is a widely used risk measure in finance that estimates potential losses over a specified time horizon at a given confidence level. Regulators include VaR in capital requirement frameworks (e.g. Basel, Solvency) and it is also employed as a key management tool.

11.3.1 Definition

Let X denote the (random) value of a financial position at the end of the period, and let R_0 be the invested capital in a risk-free asset. The *Value-at-Risk at level* p, denoted $\mathrm{VaR}_p(X)$, is defined as the smallest amount of capital m one needs to invest in the risk-free asset so that the probability of a shortfall is at most p:

$$\mathrm{VaR}_p(X) \;=\; \min\big\{m \,:\, \mathbb{P}\big(m\,R_0 + X < 0\big) \;\le\; p\big\}\,.$$

11.3.2 Properties

When the distribution of $\frac{X}{R_0}$ is continuous and strictly increasing, one can equivalently write:

$$\text{VaR}_p(X) \;=\; -\zeta_p,$$

where ζ_p is defined as the value such that:

$$F_{X/R_0}(-\zeta_p) = p.$$

That is,

$$\text{VaR}_p(X) = -F^{-1}_{X/R_0}(p).$$

Hence, we see that VaR_p is a quantile.

11.3.3 Guide to Finding the Quantile q_p

The quantile of order p (where $0 < p < 1$), denoted as q_p or $F_X^{-1}(p)$, is a value such that the probability of the random variable X being less than or equal to q_p is at least p. More formally, it's the smallest value x for which $F_X(x) \geq p$. Here's a step-by-step approach:

1. **Obtain the Cumulative Distribution Function (CDF):** The first step is always to determine or define the CDF, $F_X(x) = P(X \leq x)$.
   - For a **continuous** random variable with probability density function (PDF) $f_X(t)$, the CDF is $F_X(x) = \int_{-\infty}^{x} f_X(t)dt$.
   - For a **discrete** random variable that takes values $x_1 < x_2 < \ldots$, the CDF is a step function: $F_X(x) = \sum_{x_i \leq x} P(X = x_i)$.
2. **Set the CDF equal to p (or apply the definition):** The goal is to find the value q_p such that $F_X(q_p)$ relates to p.
3. **Solve for q_p based on the type of distribution:**
   - **For Continuous Distributions:** The quantile q_p is typically found by solving the equation:

     $$F_X(q_p) = p$$

     This often involves setting up and solving an integral:

     $$\int_{L}^{q_p} f_X(t)dt = p$$

     where L is the lower bound of the support of X. *Intuition:* You are finding the value q_p on the x-axis such that the total accumulated probability (area under the PDF) to the left of q_p is exactly p.

- **For Discrete Distributions:** The CDF is a step function. The quantile q_p is defined as the smallest value x such that $F_X(x) \geq p$.

$$q_p = \min\{x \mid F_X(x) \geq p\}$$

Intuition:

(a) List the possible values x_i of X in increasing order and their corresponding cumulative probabilities $F_X(x_i)$.

(b) Find the smallest x_i for which its cumulative probability $F_X(x_i)$ is greater than or equal to p. This x_i is your quantile q_p.

(c) Visually, if you draw the step function CDF, and draw a horizontal line at height p, q_p is the x-value where this line first hits or crosses a step of the CDF. The quantile q_p will always be one of the actual values that X can take.

In essence, the quantile q_p splits the distribution such that a proportion p of the probability mass lies to its left (or at q_p itself).

11.3.4 Interpretation

Economically, $\text{VaR}_p(X)$ represents the *minimum* capital that must be held in the risk-free asset to ensure that the total position $(mR_0 + X)$ remains non-negative with probability at least $1 - p$. If $X \geq 0$ almost surely, the total position is always positive, and thus $\text{VaR}_p(X)$ will be negative—indicating that no additional capital is required to cover risk.

11.3.5 Sign Convention

There is no universal rule about whether VaR should be reported as a positive or negative number. Commonly, VaR is presented as a *positive* value denoting the size of a potential loss, and we adopt that convention here. Note that some texts use a probability p near 1 and describe the loss distribution rather than the profit distribution; the resulting definitions are equivalent if one keeps track of signs carefully.

11.3.6 Illustration

The figure below schematically shows the probability density function of the profit–loss (P&L) distribution, with VaR_p corresponding to a lower tail quantile of order p. The left-most shaded region (of width $p \times 100\%$) represents outcomes that yield losses at least as large (in absolute value) as the VaR.

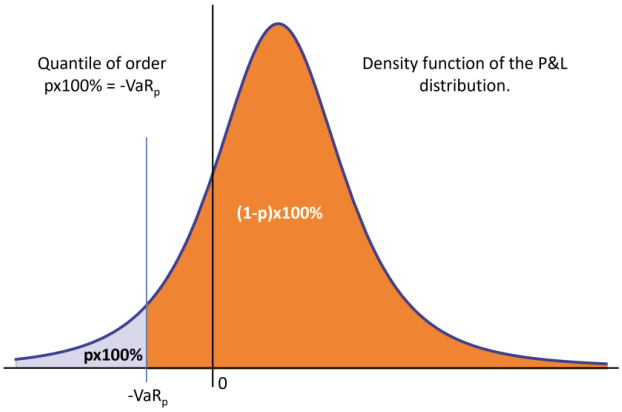


Figure 5: Density function of the P&L distribution indicating the VaR_p tail (left side) of size $p \times 100\%$.

11.3.7 Alternative Representation Using Losses

In industries such as insurance, it can be more convenient to work directly with the loss distribution rather than the value of a position. Let X be the value of our position at time 1. Define the (discounted) loss as:

$$L = -\frac{X}{R_0}.$$

Then the probability relation

$$\mathbb{P}\left(X/R_0 < -\zeta_p\right) \leq p \quad \Longleftrightarrow \quad \mathbb{P}\left(-X/R_0 > \zeta_p\right) \leq p \quad \Longleftrightarrow \quad \mathbb{P}\left(L \leq \zeta_p\right) \geq 1-p.$$

This leads to the following equivalent definition of value-at-risk in terms of the loss variable L:

$$\mathrm{VaR}_p(X) = \min\left\{m : \mathbb{P}(L \leq m) \geq 1-p\right\}.$$

When the loss distribution F_L is continuous and strictly increasing, this becomes:

$$F_L(\mathrm{VaR}_p(X)) = 1-p \quad \Longleftrightarrow \quad \mathrm{VaR}_p(X) = F_L^{-1}(1-p).$$

11.3.8 Interpretation and Notation Variants

- The quantity $\mathrm{VaR}_p(X)$ can be interpreted as the smallest value such that the losses L will be smaller than this value with probability at least $1-p$.
- Equivalently, $\mathrm{VaR}_p(X)$ can be seen as the minimal amount which, if invested in the risk-free asset, ensures that the total losses at the end of the period are covered with probability at least $1-p$.

- The value-at-risk formulation may vary with context. When working in terms of losses, we may denote the risk level as $\mathrm{VaR}_{1-p}(L)$, since:

$$L = -\frac{X}{R_0},$$

and hence, in the continuous case,

$$\mathrm{VaR}_p(X) = -F^{-1}_{X/R_0}(p) = F^{-1}_{-X/R_0}(1-p) = F^{-1}_{L}(1-p) = \mathrm{VaR}_{1-p}(L).$$

11.3.9 Illustration: Value-at-Risk from the Loss Distribution

The figure below shows the probability density function of the loss variable

$$L = -\frac{X}{R_0},$$

and visually represents how the value-at-risk VaR_p separates the left $1-p$ proportion of the distribution (where losses are acceptable) from the right p proportion (where losses exceed the risk threshold).

The shaded area under the curve corresponds to the probability mass in the right tail of the distribution, equal to $p \times 100\%$. This is the probability that losses exceed the value-at-risk threshold, and hence the capital level defined by VaR_p ensures that such losses occur with at most probability p.

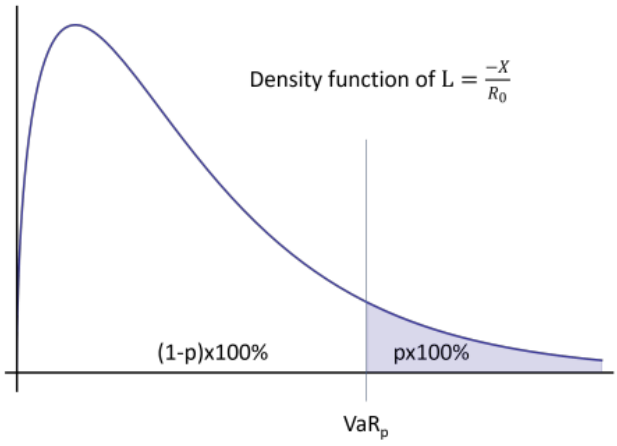


Figure 6: Value-at-risk VaR_p illustrated on the loss distribution $L = -X/R_0$.

11.3.10 Properties of Value-at-Risk

Axiomatic Properties. The Value-at-Risk measure $\mathrm{VaR}_p(X)$ satisfies the following standard properties of a risk measure:

- **Cash invariance:** $\mathrm{VaR}_p(X + c) = \mathrm{VaR}_p(X) - c$, for any constant $c \in \mathbb{R}$. Adding cash to a position reduces risk by the same amount.

- **Monotonicity:** If $X \leq Y$ almost surely, then $\mathrm{VaR}_p(X) \geq \mathrm{VaR}_p(Y)$. A position with worse outcomes is riskier.
- **Positive homogeneity:** For any $\lambda \geq 0$, $\mathrm{VaR}_p(\lambda X) = \lambda \,\mathrm{VaR}_p(X)$. Scaling a position scales its risk proportionally.

Limitations. Despite satisfying the properties above, Value-at-Risk does not satisfy subadditivity in general, which limits its coherence as a risk measure:

- **No guarantee of diversification benefits:** VaR_p does not necessarily reward diversification. That is, it is possible that

$$\mathrm{VaR}_p(X+Y) > \mathrm{VaR}_p(X) + \mathrm{VaR}_p(Y),$$

  even when X and Y are independent or identically distributed.
- **Not subadditive in general:** Subadditivity is a key axiom of coherent risk measures. The lack of this property means VaR_p may underestimate the benefit of risk pooling and can discourage diversification in portfolio management.

11.3.11 Example: Lack of Subadditivity of Value-at-Risk

To illustrate that VaR_p is not necessarily subadditive, consider a simple discrete setting involving an insurance company exposed to two sources of risk. Let the portfolio loss $L = -X$ be a discrete random variable with cumulative distribution function

$$F_L(x) = \begin{cases} 0, & x < 1, \\ 0.2, & 1 \leq x < 2, \\ 0.95, & 2 \leq x < 4, \\ 0.99, & 4 \leq x < 5, \\ 1, & x \geq 5. \end{cases}$$

We compute the value-at-risk at confidence levels 0.95 and 0.99:

$$\mathrm{VaR}_{0.95}(L) = 2, \quad \mathrm{VaR}_{0.99}(L) = 4.$$

Disaggregated Losses. Assume now that the portfolio consists of two separate policies, with losses Z and Y having cumulative distributions:

$$F_Z(x) = \begin{cases} 0, & x < 0, \\ 0.01, & 0 \leq x < 1, \\ 0.21, & 1 \leq x < 2, \\ 0.96, & 2 \leq x < 4, \\ 1, & x \geq 4, \end{cases} \qquad F_Y(x) = \begin{cases} 0, & x < 0, \\ 0.99, & 0 \leq x < 5, \\ 1, & x \geq 5. \end{cases}$$

Then we compute:

$$\mathrm{VaR}_{0.955}(Z) = 2, \quad \mathrm{VaR}_{0.955}(Y) = 0.$$

So for the combined loss $L = Z + Y$, we get:

$$\mathrm{VaR}_{0.955}(L) = 4 > \mathrm{VaR}_{0.955}(Z) + \mathrm{VaR}_{0.955}(Y) = 2.$$

This clearly violates the subadditivity condition:

$$\mathrm{VaR}_{0.955}(L) \not\leq \mathrm{VaR}_{0.955}(Z) + \mathrm{VaR}_{0.955}(Y).$$

Conclusion. We conclude that VaR_p is not a subadditive risk measure, even in the case of independent and identically distributed components.

Constructing the Decomposition. To complete the example, define the losses Z and Y as:

$$Z = \begin{cases} 1, & L \leq 4, \\ 0, & \text{otherwise}, \end{cases} \qquad Y = \begin{cases} 0, & L \leq 4, \\ 4, & L > 4. \end{cases}$$

Then clearly $Z + Y = L$, which verifies the construction and completes the counterexample.

11.3.12 VaR Applied to Normally Distributed Returns

Assume the daily returns of an asset are normally distributed with mean μ and variance σ^2. Let returns over T days be independent and identically distributed (i.i.d.), and suppose $R_0 = 1$.

Single Asset Case. Let r_T be the return over T days, then:

$$r_T = \mu T + \sigma\sqrt{T}Z, \quad \text{where } Z \sim \mathcal{N}(0,1).$$

The value-at-risk is given by:

$$\mathrm{VaR}_p(r_T) = -F_{r_T}^{-1}(p) = -\left(\sigma\sqrt{T}\Phi^{-1}(p) + \mu T\right) = \sigma\sqrt{T}\Phi^{-1}(1-p) - \mu T,$$

where $\Phi^{-1}(p)$ denotes the quantile function (inverse CDF) of the standard normal distribution.

Two-Asset Portfolio. Now consider a portfolio with two assets, with returns:

$$r_{i,T} \sim \mathcal{N}(\mu_i T, \sqrt{T}\sigma_i), \quad i = 1, 2,$$

not necessarily i.i.d.

- The sum of individual VaRs is:

$$\mathrm{VaR}_p(r_{1,T}) + \mathrm{VaR}_p(r_{2,T}) = -\left((\sigma_1 + \sigma_2)\sqrt{T}\Phi^{-1}(p) + (\mu_1 + \mu_2)T\right).$$

- The VaR of the portfolio (with total return $r_{1,T} + r_{2,T}$) is:

$$\mathrm{VaR}_p(r_{1,T}+r_{2,T}) = -\left(\sqrt{\sigma_1^2 + \sigma_2^2 + 2\,\mathrm{corr}(r_1, r_2)\sigma_1\sigma_2} \cdot \sqrt{T}\Phi^{-1}(p) + (\mu_1 + \mu_2)T\right).$$

Note: The formula uses the general variance of a sum:

$$\mathrm{Var}(w_1 r_1 + w_2 r_2) = w_1^2\sigma_1^2 + w_2^2\sigma_2^2 + 2w_1 w_2\,\mathrm{Cov}(r_1, r_2).$$

In the symmetric case shown, $w_1 = w_2 = 1$.

Diversification Effect. Since for common values of p (e.g., $p = 0.01, 0.05$), the quantile $\Phi^{-1}(p)$ is negative, it follows that:

$$\mathrm{VaR}_p(r_{1,T} + r_{2,T}) \leq \mathrm{VaR}_p(r_{1,T}) + \mathrm{VaR}_p(r_{2,T}).$$

Property. For normally distributed returns and typical values of p, VaR *does* reward diversification:

$$\mathrm{VaR}_p(r_1 + r_2 + \cdots + r_n) \leq \sum_{i=1}^{n} \mathrm{VaR}_p(r_i).$$

This shows that subadditivity can hold in the Gaussian case, in contrast to the general setting.

11.3.13 VaR Applied to a Lognormal Distribution

Consider an asset with initial value S_0, and let S_1 be its value at time 1. Assume the logarithmic return is normally distributed:

$$\ln\left(\frac{S_1}{S_0}\right) \sim \mathcal{N}(\mu, \sigma).$$

This implies that S_1 follows a lognormal distribution. The profit/loss from the position is $S_1 - S_0$. The value-at-risk at level p is then:

$$\mathrm{VaR}_p(S_1 - S_0) = -F^{-1}_{S_1 - S_0}(p) = F^{-1}_{S_0 - S_1}(1-p) = S_0\left(1 - e^{\mu + \sigma\Phi^{-1}(p)}\right),$$

where $\Phi^{-1}(p)$ is the quantile function of the standard normal distribution.

Interpretation.

- The distribution of $S_1 - S_0$ is right-skewed due to the lognormal nature of S_1.
- The VaR corresponds to a left-tail quantile (loss quantile) of this skewed distribution.
- An equivalent perspective is to analyze $S_0 - S_1$, which flips the distribution and centers the analysis on losses.
- The shaded region in the figure indicates the probability p of observing a loss at least as large as VaR_p.

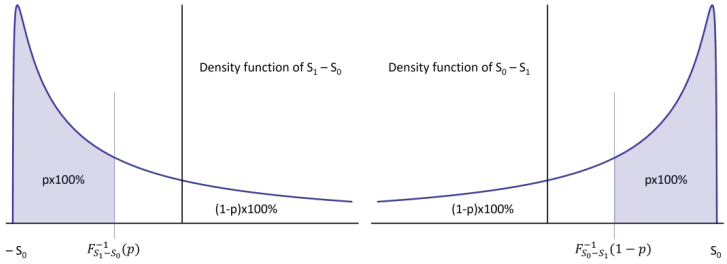


Figure 7: Value-at-risk for lognormal returns. Left: distribution of $S_1 - S_0$. Right: distribution of $S_0 - S_1$.

11.3.14 Disadvantages of Value-at-Risk

Although Value-at-Risk (VaR) is one of the most widely used and discussed risk measures in financial risk management, it has several notable drawbacks:

1. Sensitivity to Parameter Choices

- The definition of VaR requires specifying a time horizon, which can significantly influence the result.
- The confidence level p must also be chosen, and it may differ depending on the context (e.g., regulatory vs. internal risk management).

2. Lack of Subadditivity VaR may fail to satisfy subadditivity, a key property of coherent risk measures:

$$\text{VaR}_p(X + Y) > \text{VaR}_p(X) + \text{VaR}_p(Y),$$

even when X and Y are independent. This has two important consequences:

- Firms might be discouraged from diversifying, as splitting positions can appear more "capital-efficient" under VaR.

- Portfolio risk may be underestimated due to the failure to properly account for diversification effects.

3. No Information About Tail Losses

- VaR only measures the threshold loss exceeded with probability p, but does not describe how large those losses could be.
- It offers no insight into the distribution of losses beyond the VaR cutoff.

Visual Illustration: VaR Ignores Tail Severity Consider two positions with different profit and loss (P&L) distributions. The figure below shows:

- The P&L distributions of two positions (green and blue curves).
- The area in the left tail (shaded orange and blue) corresponds to probability p, i.e., the 5% or 1% worst-case outcomes.
- The value-at-risk is the same for both portfolios: $\text{VaR}_p^1 = \text{VaR}_p^2$.

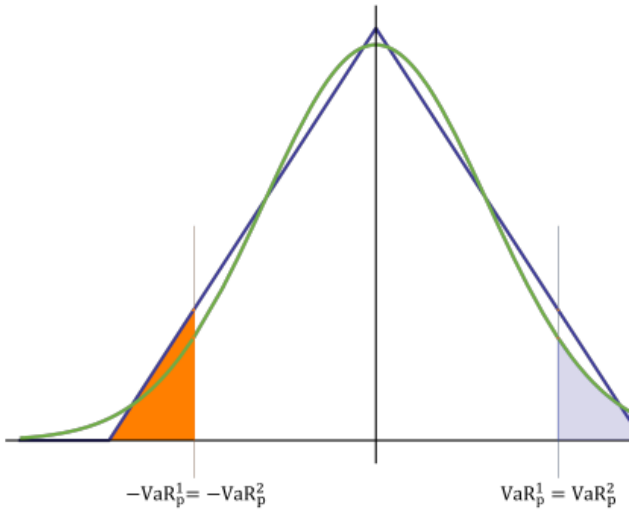


Figure 8: Two distributions with the same VaR_p but different tail risks.

However, as the second panel shows:

- The losses beyond VaR can differ dramatically.
- Portfolio 1 has a fatter tail than Portfolio 2 meaning that when losses exceed VaR, they may be much more severe.
- VaR does not capture this difference, potentially masking substantial risk.

4. Lessons from the Financial Crisis

- The 2007–2008 financial crisis highlighted severe weaknesses in the use of VaR.
- Many large financial institutions estimated their VaR in the range of hundreds of millions of dollars yet they experienced losses in the billions.
- These failures were due, in part, to overreliance on VaR and the lack of consideration for extreme tail events.

5. Regulatory Shift Toward Expected Shortfall

- In response to these limitations, regulatory bodies such as the Basel Committee on Banking Supervision (BCBS) introduced reforms under the Fundamental Review of the Trading Book (FRTB).
- Value-at-Risk is being (or has been) replaced by *Expected Shortfall* (ES), a risk measure that captures both the likelihood and severity of tail losses.

11.4 Expected Shortfall

11.4.1 Definition and Properties

One of the major weaknesses of Value-at-Risk (VaR) is that it does not provide any information about the magnitude of losses beyond the VaR threshold. To address this, one can use the *Expected Shortfall* (ES), which captures the expected loss in the tail of the loss distribution.

Expected shortfall at level p is defined as the average VaR for all levels $u \leq p$:

$$\mathrm{ES}_p(X) = \frac{1}{p} \int_0^p \mathrm{VaR}_u(X)\, du.$$

Expected shortfall is also known under various names in the literature, including:

- Average VaR,
- Conditional VaR (CVaR),
- Tail VaR,
- Tail Conditional Expectation.

Alternative Representation. There are multiple equivalent ways to define Expected Shortfall, depending on the context. A particularly intuitive one especially when dealing with continuous distributions is the conditional expectation representation:

If X is a continuous random variable, then the expected shortfall at level p satisfies:

$$\mathrm{ES}_p(X) = \mathbb{E}[L \mid L > \mathrm{VaR}_p(X)],$$

where L is the loss.

Interpretation. This representation shows that $\mathrm{ES}_p(X)$ is the expected loss given that the loss exceeds the value-at-risk threshold. That is, ES captures the *average of the worst-case losses*, whereas VaR only marks the point where tail losses begin.

Property. Expected shortfall is a **coherent risk measure**, satisfying:

- *Monotonicity*
- *Cash invariance*
- *Positive homogeneity*
- *Subadditivity*

This makes ES a more robust and theoretically sound alternative to VaR, particularly in the presence of heavy-tailed loss distributions or when managing portfolio risk.

11.4.2 ES Applied to Normally Distributed Returns

Let the daily returns r_1 of an asset be normally distributed with mean μ and standard deviation σ, i.e.

$$r_1 = \mu + \sigma Z, \quad \text{where } Z \sim \mathcal{N}(0,1),$$

and assume $R_0 = 1$.

The value-at-risk at confidence level p is given by:

$$\mathrm{VaR}_p(r_1) = -F_{r_1}^{-1}(p) = -\left(\sigma\Phi^{-1}(p) + \mu\right) = \sigma\Phi^{-1}(1-p) - \mu,$$

where $\Phi^{-1}(p)$ is the quantile of the standard normal distribution.

Expected Shortfall Formula. To compute the expected shortfall $\mathrm{ES}_p(r_1)$, one integrates the VaR over all levels $u \le p$:

$$\mathrm{ES}_p(r_1) = \frac{1}{p}\int_0^p \mathrm{VaR}_u(r_1)\,du = \frac{1}{p}\int_0^p \left(\sigma\Phi^{-1}(1-u) - \mu\right) du.$$

This simplifies to:

$$\mathrm{ES}_p(r_1) = \sigma\frac{\phi(\Phi^{-1}(1-p))}{p} - \mu,$$

where ϕ denotes the probability density function of the standard normal distribution.

Interpretation. This result shows that:

- The expected shortfall exceeds the value-at-risk, capturing the severity of tail losses.
- As p decreases (i.e., as we consider more extreme scenarios), the ratio $\phi(\Phi^{-1}(1-p))/p$ increases, leading to a higher ES.
- For normally distributed returns, ES has a closed-form expression that depends only on μ, σ, and p.

Note. This explicit formula for ES under normality is convenient in practice, and makes it easier to compute both portfolio and asset-level expected shortfall using analytical methods.

11.4.3 Expected Shortfall for Student's t-Distributed Returns

Let the daily return r_1 of an asset be modeled as:

$$r_1 = \mu + \sigma Z,$$

where Z follows a Student's t-distribution with ν degrees of freedom: $Z \sim t_\nu$.

Properties of the Student's t-Distribution:

- $\mathbb{E}[Z] = 0$ and $\mathrm{Var}(Z) = \frac{\nu}{\nu-2}$, for $\nu > 2$.
- The distribution is symmetric around 0, similar to the standard normal distribution.
- The probability density function (pdf) is given by:

$$g_\nu(x) = \frac{\Gamma((\nu+1)/2)}{\sqrt{\nu\pi}\,\Gamma(\nu/2)} \left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}.$$

Value-at-Risk under the t-Distribution. As with the normal case, the VaR of r_1 is:

$$\mathrm{VaR}_p(r_1) = \sigma t_\nu^{-1}(1-p) - \mu,$$

where $t_\nu^{-1}(1-p)$ is the quantile of the Student's t-distribution with ν degrees of freedom.

Expected Shortfall under the t-Distribution. The expected shortfall is given by:

$$\mathrm{ES}_p(r_1) = \sigma \frac{g_\nu(t_\nu^{-1}(1-p))}{p} \left(\frac{\nu + (t_\nu^{-1}(1-p))^2}{\nu - 1}\right) - \mu,$$

where:

- $g_\nu(x)$ is the density of the t-distribution,
- $t_\nu^{-1}(1-p)$ is the $1-p$ quantile,
- $\nu > 1$ for the expectation to be finite.

11.4.4 Comparison: Normal vs. Student's t-Distribution

- Both distributions are symmetric and commonly used for modeling log-returns.
- The t-distribution has heavier tails than the normal distribution.
- Heavier tails imply a higher likelihood of extreme events this makes the t-distribution more realistic for risk modeling, especially in stressed markets.
- As $\nu \to \infty$, the t-distribution converges to the normal distribution.

11.5 Empirical Methods

Empirical methods are widely used in risk management to estimate future values based on past data. The approach relies on several key assumptions:

- Historical data (such as prices or other quantities) are used to model the future values of a portfolio.
- It is assumed that the information contained in the historical sample is representative of future values.
- The mechanism that generated past returns is assumed to remain consistent over time.
- This is a subjective, fully non-parametric method, yet it is still considered a reasonable approach in many practical applications.

11.5.1 Sample Preparation

Transforming prices into percentage returns is not mandatory for sample preparation to be effective; however, it is often convenient.

- Let $\{\mathbf{S}_{-n}, \mathbf{S}_{-n+1}, \ldots, \mathbf{S}_0\}$ be the initial historical sample of asset prices.
- From this, we derive a new sample $\{\mathbf{Z}_{-n+1}, \mathbf{Z}_{-n+2}, \ldots, \mathbf{Z}_0\}$, which can be viewed as sampled from the distribution of V_1, the portfolio value at time 1.

Notation and Definitions:

- V_1 is the random portfolio value at time 1.
- $\mathbf{S}_1$ is a d-dimensional vector of asset prices at time 1.
- $\mathbf{R}_1$ is the vector of percentage returns in period 1.
- $V_1 = f_1(\mathbf{S}_1) = f_2(\mathbf{R}_1)$, where f_1 and f_2 depend only on initial data.
- Sample of historical prices: $\{\mathbf{S}_{-n}, \ldots, \mathbf{S}_0\}$ with $\mathbf{S}_{-k} = (S^1_{-k}, S^2_{-k}, \ldots, S^d_{-k})$.
- Time between observations is equal.

Return Vectors:

- Prices today are not necessarily representative of future prices.
- Instead, we transform the prices into returns: $\mathbf{R}_{-k} = (R^1_{-k}, \ldots, R^d_{-k})^T$, with $R^l_{-k} = \frac{S^l_{-k}}{S^l_{-k-1}}$.
- $\{\mathbf{R}_{-n+1}, \ldots, \mathbf{R}_0\}$ is assumed to be a sample of $\mathbf{R}_1$, and consequently $\{f_2(\mathbf{R}_{-n+1}), \ldots, f_2(\mathbf{R}_0)\}$ is a sample of V_1.

Example:

- Consider a portfolio with long positions in 3 assets, with holdings $i \in \{1, 2, 3\}$.
- Then,

$$V_1 = S^1_1 + 2S^2_1 + 3S^3_1 = S^1_0 R^1_1 + 2S^2_0 R^2_1 + 3S^3_0 R^3_1 = f_2(\mathbf{R}_1)$$

- We can use the last 30 days of returns, i.e., the sample $\{\mathbf{R}_{-29}, \ldots, \mathbf{R}_0\}$.

11.5.2 Obtaining Future Samples V_t, $t > 1$:

To estimate values at time $t > 1$, different methods are used:

Thinning of the Sample: The goal of thinning is to generate a sample from the distribution of V_t, the portfolio value at time t, using historical return data. This is done by constructing non-overlapping t-period returns from the original historical sample.

- Assume we have access to a historical sample of asset prices:

$$\{\mathbf{S}_{-n}, \mathbf{S}_{-n+1}, \ldots, \mathbf{S}_0\}$$

- Define the portfolio value at time t as:

$$V_t = f(\mathbf{R}_1, \mathbf{R}_2, \ldots, \mathbf{R}_t) = f(\mathbf{R}_1 \cdot \mathbf{R}_2 \cdots \mathbf{R}_t)$$

  where the product is understood componentwise.

- We now form a thinned sample of prices using a step size of t, selecting every t-th observation:

$$\{\mathbf{S}_{-t\lfloor n/t \rfloor}, \ldots, \mathbf{S}_{-2t}, \mathbf{S}_{-t}, \mathbf{S}_0\}$$

  where $\lfloor n/t \rfloor$ is the integer part of n/t, i.e., the number of t-length segments that fit in the historical window.

- Using these selected price vectors, we define return vectors over t periods as:

$$\mathbf{R}_k^T = \left((R_k^1)^{(t)}, \ldots, (R_k^d)^{(t)}\right)^T$$

  where each component is computed as:

$$(R_k^l)^{(t)} = \frac{S_k^l}{S_{k+t}^l}, \quad l = 1, \ldots, d$$

- In matrix notation, we can express the sample of t-period return vectors as:

$$\{\mathbf{R}_{\lfloor n/t \rfloor + 1}^T, \ldots, \mathbf{R}_1^T, \mathbf{R}_0^T\}$$

- Applying the portfolio value function f to each of these return vectors gives a sample from the distribution of V_t:

$$\{f(\mathbf{R}_{\lfloor n/t \rfloor + 1}^T), \ldots, f(\mathbf{R}_1^T), f(\mathbf{R}_0^T)\}$$

Remarks and Limitations:

- The resulting t-period returns can be viewed as approximately independent and identically distributed (i.i.d.) copies of $\mathbf{R}_1 \cdots \mathbf{R}_t$.

- However, the main downside is a reduced sample size, since we only use every t-th observation from the price sample.

- Additionally, some information contained in the original price sample is discarded, as overlapping data is ignored.

Historical Simulation: Historical simulation offers an alternative way to obtain a sample from the distribution of V_t by reusing observed return data in a randomized, resampled fashion. This approach does not rely on thinning but instead leverages bootstrapping.

- Let $\{\mathbf{R}_{-n+1}, \ldots, \mathbf{R}_0\}$ be the historical return sample.

- We aim to simulate t-period returns. To do this, we draw t return vectors *with replacement* from the historical return sample.
- Denote each resampled t-length path as:

$$(\mathbf{R}_{i_1}, \mathbf{R}_{i_2}, \ldots, \mathbf{R}_{i_t})$$

  where each $i_j \in \{-n+1, \ldots, 0\}$ is randomly selected with replacement.
- Compute the componentwise product of these t return vectors to obtain the synthetic t-period return:

$$\mathbf{R}_1^{*(t)} = \mathbf{R}_{i_1} \cdot \mathbf{R}_{i_2} \cdots \mathbf{R}_{i_t}$$

- Repeat this sampling process m times to generate a set of m fictive t-period return vectors:

$$\left\{\mathbf{R}_1^{*(t)}, \ldots, \mathbf{R}_m^{*(t)}\right\}$$

- Then, by applying the portfolio value function f, we obtain a sample from the distribution of V_t:

$$\left\{f(\mathbf{R}_1^{*(t)}), \ldots, f(\mathbf{R}_m^{*(t)})\right\}$$

Advantages and Disadvantages:

- **Advantage:** All available historical return vectors are utilized in the resampling procedure.
- **Advantage:** The method can generate an arbitrarily large sample by increasing m.
- **Disadvantage:** The resulting vectors $\mathbf{R}_1^{*(t)}, \ldots, \mathbf{R}_m^{*(t)}$ are identically distributed, but *not* independent. This is because the same return vectors may be selected multiple times.
- **Disadvantage:** Repetition of the same return vectors may lead to redundant fictive returns and limited variability in the simulated sample.

11.5.3 Empirical Distribution

To estimate risk measures such as the Value-at-Risk (VaR) and Expected Shortfall (ES), we need to evaluate the underlying distribution function of portfolio values. In empirical methods, we adopt a non-parametric approach and estimate this distribution using the empirical cumulative distribution function (ECDF).

Definition: Let $X_1, \ldots, X_n$ be independent and identically distributed (i.i.d.) random variables with corresponding observed values $x_1, \ldots, x_n$.

- The true cumulative distribution function (CDF) of a random variable X (an i.i.d. copy of the X_k) is defined as:

$$F_X(x) = \mathbb{P}(X \leq x) = \mathbb{P}(X_1 \leq x_1, \ldots, X_d \leq x_d)$$

- Since $F_X(x)$ is generally unknown, we approximate it using the empirical distribution function:

$$F_n(x) = \frac{1}{n} \sum_{k=1}^{n} \mathbb{I}\{x_k \leq x\}$$

  where $\mathbb{I}\{\cdot\}$ is the indicator function that takes the value 1 if the condition is true, and 0 otherwise.

Properties:

- $F_n(x)$ is a function in $\mathbb{R}$ and is known as the empirical distribution function based on the sample $x_1, \ldots, x_n$.
- If the observations themselves are random (i.e., we work with $X_1, \ldots, X_n$), the empirical distribution function becomes a random function:

$$F_{n,X}(x) = \frac{1}{n} \sum_{k=1}^{n} \mathbb{I}\{X_k \leq x\}$$

- This random function converges to the true distribution function $F_X(x)$ almost surely as $n \to \infty$.

Convergence and the Law of Large Numbers: By the Strong Law of Large Numbers (SLLN), we have the following results:

$$F_{n,X}(x) \to F_X(x) \quad \text{as } n \to \infty \text{ with probability 1}$$
$$\frac{1}{n} \sum_{k=1}^{n} h(X_k) \to \mathbb{E}[h(X)] \quad \text{as } n \to \infty \text{ with probability 1}$$

where h is any integrable function. This justifies the use of empirical distributions in risk management, as they provide consistent estimators for the true underlying distribution and its moments.

11.5.4 Empirical Quantiles

Motivation and Intuition From the empirical distribution, we can estimate the relevant quantile (or Value-at-Risk). This estimated value is what we call the *empirical quantile*.

Final Form of the Empirical Quantile The empirical quantile—whether random or observed—boils down to selecting the appropriate order statistic in the sample:

- For the **random empirical quantile** (when the sample $\{X_1, \ldots, X_n\}$ is itself random), we have:

$$F_{n,X}^{-1}(p) = X_{1+\lfloor n(1-p)\rfloor,n},$$

  where $X_{j,n}$ is the j-th order statistic in the sample (ordered from largest to smallest, in this convention).

- For the **observed empirical quantile** (once you have a realized dataset $\{x_1, \ldots, x_n\}$), you similarly get:

$$F_n^{-1}(p) = x_{1+\lfloor n(1-p)\rfloor,n},$$

  where $x_{j,n}$ is the j-th order statistic from the actual observed data (again, from largest to smallest).

Asymptotic Consistency One of the reasons empirical quantiles are so important is that, under mild conditions, they *consistently* estimate the true quantiles. Specifically, if the true distribution function F is continuous and strictly increasing, then for each fixed $p \in (0, 1)$,

$$\mathbb{P}\Big(\lim_{n\to\infty} F_{n,X}^{-1}(p) = F^{-1}(p)\Big) \;=\; 1.$$

This says that as the sample size n grows large, the empirical p-quantile *converges almost surely* to the true quantile $F^{-1}(p)$. In practice, this justifies using empirical quantiles (e.g., for Value-at-Risk calculation) when sufficient data is available.
Key takeaway: The empirical quantile is a straightforward but powerful idea—take your sample, sort it, and pick the value whose position in the sorted list corresponds to the fraction p. In large samples and under decent assumptions, this value is a reliable estimator of the true quantile from which the data were drawn.

11.5.5 Empirical VaR and ES

In risk management, the *loss variable* L is often defined in terms of the portfolio value. If X is the value of the portfolio at time 1, and X_0 is the initial value, then the loss can be defined as:

$$L = -\frac{X}{X_0}$$

Let $L_1, \ldots, L_n$ be an i.i.d. sample from the distribution of L, and let $L_{1,n} \geq L_{2,n} \geq \cdots \geq L_{n,n}$ denote the ordered statistics.

Empirical Value-at-Risk (VaR): The empirical Value-at-Risk at level $p \in (0,1)$, denoted $\widehat{\mathrm{VaR}}_p$, is the empirical quantile of order $1-p$. It is computed as:

$$\widehat{\mathrm{VaR}}_p(X) = L_{1+\lfloor np \rfloor, n}$$

Empirical Expected Shortfall (ES): The Expected Shortfall (ES), also known as Conditional VaR, measures the expected loss given that the loss exceeds the VaR threshold. The empirical version is given by:

$$\widehat{\mathrm{ES}}_p = \frac{1}{p} \int_0^p L_{1+\lfloor nu \rfloor, n} \, du$$

This integral can be approximated as a weighted sum:

$$\widehat{\mathrm{ES}}_p = \frac{1}{p} \left(\sum_{i=1}^{\lfloor np \rfloor} \frac{L_{i,n}}{n} + \left(p - \frac{\lfloor np \rfloor}{n} \right) L_{1+\lfloor np \rfloor, n} \right)$$

This formula accounts for the discrete nature of the ordered sample and interpolates to estimate the expected loss beyond the empirical VaR.

11.5.6 Confidence Intervals

Definition Let θ be the true but unknown quantity of interest—here, θ could be the true quantile $F^{-1}(p)$ of the loss distribution F. A $(1-\alpha)$-confidence interval (A, B) for θ is defined so that

$$\mathbb{P}\big(A \le \theta \le B\big) \;=\; 1 - \alpha,$$

where $1-\alpha \in (0,1)$ is the desired confidence level (often denoted by q in some texts). For simplicity, we re-write to A and B such that

$$\mathbb{P}\big(F^{-1}(p) \le A\big) \;=\; \mathbb{P}\big(F^{-1}(p) \ge B\big) \;=\; \frac{\alpha}{2},$$

giving a total coverage of $1-\alpha$. In practice, because of finite sample sizes and discrete order statistics, you may not get an *exactly* centered interval, but the idea is to have roughly equal tails.

Using Order Statistics to Bracket $F^{-1}(p)$. Because F is unknown, we cannot directly solve $\mathbb{P}(A \le F^{-1}(p) \le B) = 1-\alpha$. Instead, we *approximate* F by its empirical distribution and use *order statistics*.

We then want indices i and j so that, with high probability, the true quantile $F^{-1}(p)$ lies between $X_{i,n}$ and $X_{j,n}$.

$$\mathbb{P}\big(X_{i,n} < F_X^{-1}(p) < X_{j,n}\big) \;=\; q',$$

where q' is as close as possible to the target confidence level $(1-\alpha)$. Often, we split the probability so that

$$\mathbb{P}\big(F_X^{-1}(p) \le X_{i,n}\big) \;\approx\; \mathbb{P}\big(F_X^{-1}(p) \ge X_{j,n}\big) \;\approx\; \frac{1-q'}{2}.$$

Binomial Interpretation via $Y_{F_X^{-1}(p)}$ To make the coverage calculation clearer, define the random variable

$$Y_{F_X^{-1}(p)} \;=\; \sum_{k=1}^{n} \mathbb{I}\big\{X_k \le F_X^{-1}(p)\big\}.$$

This counts how many of the n observed data points $X_1, X_2, \ldots, X_n$ lie at or below the *true* quantile $F_X^{-1}(p)$.

We have that $Y_{F_X^{-1}(p)}$ follow a binomial distribution:

Each indicator $\mathbb{I}\{X_k \le F_X^{-1}(p)\}$ is like a *Bernoulli trial* with success probability p. Indeed, under our assumption that $X_1, \ldots, X_n$ are i.i.d. from a continuous distribution F, we have:

$$\mathbb{P}(X_k \le F_X^{-1}(p)) \;=\; F\big(F_X^{-1}(p)\big) \;=\; p.$$

Hence, each indicator is 1 with probability p and 0 with probability $1-p$. Since there are n independent observations, the sum of these Bernoulli variables must follow a Binomial(n, p) distribution. In other words,

$$Y_{F_X^{-1}(p)} \;\sim\; \text{Binomial}(n,\, p).$$

Consequently, any statement about "how many sample points lie below $F_X^{-1}(p)$" can be turned into a binomial probability statement involving $Y_{F_X^{-1}(p)}$.

Finding Indices i and j via the Binomial Distribution Our goal is to bracket the true quantile $F_X^{-1}(p)$ between two order statistics, $X_{i,n}$ and $X_{j,n}$, with high probability (e.g., $1-\alpha$).

We then choose i and j by referring to the tails of the Binomial(n, p) distribution. Specifically, i and j are chosen such that

$$\mathbb{P}\big(Y_{F_X^{-1}(p)} \le i-1\big) \;\approx\; \frac{\alpha}{2} \quad \text{and} \quad \mathbb{P}\big(Y_{F_X^{-1}(p)} \ge j\big) \;\approx\; \frac{\alpha}{2}.$$

The index $i-1$ effectively corresponds to the $\alpha/2$ quantile of the Binomial(n, p) distribution (i.e., we seek the largest integer $k = i-1$ such that $\mathbb{P}(Y_{F_X^{-1}(p)} \le k) \approx \alpha/2$). The index j is chosen such that the probability in the upper tail, $\mathbb{P}(Y_{F_X^{-1}(p)} \ge j)$, is also approximately $\alpha/2$ (i.e., we seek the smallest integer $k = j$ such that $\mathbb{P}(Y_{F_X^{-1}(p)} \ge k) \approx \alpha/2$).

Numerical Example Imagine $n = 200$ i.i.d. losses and you want a 95% confidence interval ($\alpha = 0.05$) around the 5%-quantile $F_X^{-1}(0.05)$. Define

$$Y_{F_X^{-1}(0.05)} \;\sim\; \text{Binomial}\big(200,\, 0.05\big).$$

You look up in a binomial table (or use software) to find indices i and j such that:

$$\mathbb{P}\big(Y_{F_X^{-1}(0.05)} \leq i-1\big) \;\approx\; 0.025 \quad \text{and} \quad \mathbb{P}\big(Y_{F_X^{-1}(0.05)} \geq j\big) \;\approx\; 0.025.$$

In one typical outcome:

$$i = 5, \quad j = 17.$$

Thus,

$$\mathbb{P}\big(X_{5,200} \leq F_X^{-1}(0.05)\big) \approx 0.975, \quad \mathbb{P}\big(X_{17,200} \geq F_X^{-1}(0.05)\big) \approx 0.975.$$

Hence, $\big[X_{5,200},\, X_{17,200}\big]$ is (approximately) a 95% confidence interval for the 5%-quantile. In practice:

1. Sort the sample in increasing order.
2. The 5th order statistic, $X_{5,200}$, is the lower end of the CI.
3. The 17th order statistic, $X_{17,200}$, is the upper end of the CI.

11.6 Portfolio Optimization

How much should one invest in each asset available? This fundamental question in portfolio theory depends on:

- the investor's preferences for risk and return,
- how the investor quantifies or measures risk.

This allocation problem can be formulated as a *constrained optimization problem*, where the objective is typically to either:

- maximize expected return for a given level of risk, or
- minimize risk for a given level of expected return.

Such problems are commonly solved using **Lagrange multipliers**, which allow us to incorporate constraints (e.g., budget constraints or restrictions on asset weights) into the optimization.

Risk Measurement

There are many quantitative measures of financial risk. One of the most commonly used measures is the *variance* of returns, or equivalently, its square root: the *standard deviation*. These are standard statistical measures of dispersion.

- Variance measures the deviation of returns from the mean (historical) or from the expected return (forward-looking).
- It is easy to compute and has well-established statistical properties.

However, using variance as a risk measure has certain limitations in financial contexts:

- It treats upward and downward deviations equally, although investors are typically more concerned with losses.
- It ignores higher moments of the return distribution such as skewness and kurtosis.
- It may fail to capture investor preferences accurately in some cases (e.g., when stochastic dominance considerations apply).

Despite these limitations, we will first use **variance** as our risk measure and consider alternative measures later.

Notation

We consider an investor with an initial capital V_0. The investor can allocate this capital among n risky assets and one risk-free asset.

- Let S_k^t denote the spot price of the k-th risky asset at time t.
- Let h_k be the investor's position (number of units held) in the k-th risky asset.
- Let $\mathbf{h} = (h_1, h_2, \ldots, h_n)^\top$ be the vector of positions in risky assets.
- The investor may also invest in a risk-free bond with price B_0 at time 0 that pays 1 monetary unit at time 1.
- Let h_0 denote the position in the risk-free asset.

The investment strategy $(h_0, \mathbf{h})$ must satisfy the budget constraint:

$$h_0 B_0 + \sum_{k=1}^{n} h_k S_k^0 \leq V_0.$$

Project Value at Time 1: The value of the investment at time 1 is given by:

$$V_1 = h_0 + \sum_{k=1}^{n} h_k S_k^1.$$

To simplify notation, we define the **amount** invested in each asset at time 0:

$$w_k = h_k S_k^0, \quad w_0 = h_0 B_0.$$

Thus, the strategy can be rewritten as $(w_0, \mathbf{w})$, where:

$$\mathbf{w} = (w_1, w_2, \ldots, w_n)^\top,$$

and the budget constraint becomes:

$$w_0 + \sum_{k=1}^{n} w_k \leq V_0.$$

Return Representation: The value at time 1 in terms of returns can be written as:

$$V_1 = w_0 \cdot \frac{1}{B_0} + \sum_{k=1}^{n} w_k \cdot \frac{S_k^1}{S_k^0}.$$

Let:

- $r_k = \frac{S_k^1}{S_k^0}$ be the gross return of asset k,
- $r_0 = \frac{1}{B_0}$ be the gross return of the risk-free asset,
- $\mathbf{r} = (r_1, r_2, \ldots, r_n)^\top$.

Then the project value becomes:

$$V_1 = w_0 r_0 + \sum_{k=1}^{n} w_k r_k.$$

Expected Return and Risk: Let:

- $\boldsymbol{\mu} = \mathbb{E}[\mathbf{r}]$ denote the vector of expected returns,
- $\boldsymbol{\Sigma} = \mathrm{Cov}(\mathbf{r})$ denote the variance-covariance matrix of asset returns.

Mean-Variance Optimization Problems

In the context of portfolio optimization, an investor typically seeks to make informed trade-offs between risk and return. Three classical formulations of the mean-variance optimization problem are:

1. **Maximize the expected return** of the portfolio subject to a constraint on the portfolio's variance:

$$\max_{\mathbf{w}} \ \mathbb{E}[V_1] \quad \text{subject to} \quad \mathrm{Var}(V_1) \leq \sigma_{\max}^2$$

2. **Minimize the variance** of the portfolio subject to a minimum expected return:

$$\min_{\mathbf{w}} \ \mathrm{Var}(V_1) \quad \text{subject to} \quad \mathbb{E}[V_1] \geq \mu_{\min}$$

3. **Maximize a trade-off** between expected return and risk (variance), typically in the form of a utility function:

$$\max_{\mathbf{w}} \ \mathbb{E}[V_1] - \lambda \cdot \mathrm{Var}(V_1)$$

where $\lambda > 0$ represents the investor's risk aversion.

These problems form the foundation of the Markowitz mean-variance framework.

Extensions of the model may incorporate more realistic market features, including:

- the existence of a risk-free asset,
- liabilities or future obligations,
- transaction costs,
- different borrowing and lending rates,
- and additional market frictions.

Efficient Portfolio and Efficient Frontier An efficient portfolio is a portfolio that offers the highest possible expected return for a given level of risk, or conversely, the lowest possible level of risk for a given expected return. The efficient frontier is the set of all efficient portfolios

The Two Asset Case To gain further insight, we consider a simplified portfolio consisting of only two risky assets. Assume the total initial capital is normalized such that $V_0 = 1$. This implies that the portfolio weights must satisfy $w_1 + w_2 = 1$.

Let the expected returns of the assets be μ_1 and μ_2, and their variances be σ_1^2 and σ_2^2, respectively. Denote by ρ the correlation coefficient between the returns of the two assets. Then the expected value and variance of the portfolio at time 1 are given by:

$$\mathbb{E}[V_1] = w_1\mu_1 + w_2\mu_2, \quad \mathrm{Var}(V_1) = w_1^2\sigma_1^2 + w_2^2\sigma_2^2 + 2\rho w_1 w_2 \sigma_1\sigma_2.$$

To simplify notation, let $w = w_1$, so $w_2 = 1 - w$. We then define:

$$\mu_{V_1}(w) := \mathbb{E}[V_1] = w\mu_1 + (1-w)\mu_2,$$

$$\sigma^2_{V_1}(w) := \mathrm{Var}(V_1) = w^2\sigma_1^2 + (1-w)^2\sigma_2^2 + 2\rho w(1-w)\sigma_1\sigma_2.$$

This allows us to analyze how different portfolio weights $w \in [0,1]$ trace out combinations of risk and return in the (σ, μ)-plane, where $\sigma = \sqrt{\mathrm{Var}(V_1)}$ is the portfolio's standard deviation (i.e., risk), and $\mu = \mathbb{E}[V_1]$ is its expected return.

In the figure below, we illustrate the efficient frontier for two extreme cases:

- **Left plot:** when $\rho = -1$, the portfolio variance can be reduced to zero through perfect negative correlation. The efficient frontier appears as a V-shaped curve.
- **Right plot:** when $\rho = 1$, the portfolio variance is a convex combination of individual asset variances without any diversification benefit. The frontier is a straight line between the two assets.

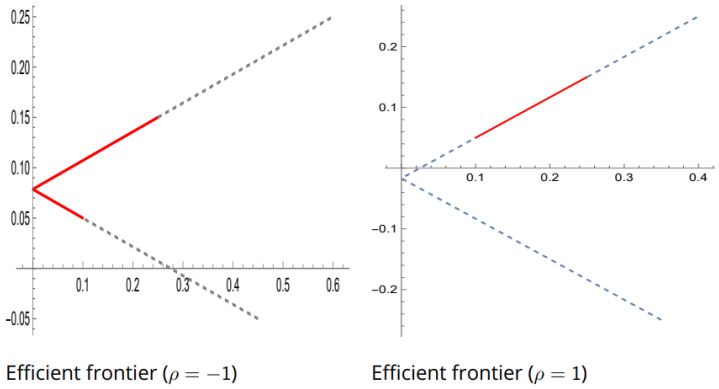


Figure 9: Efficient frontiers in the two-asset case. The x-axis represents portfolio risk (standard deviation σ), and the y-axis represents expected return μ.

Effect of Correlation and Short-Selling In the two-asset case, the shape of the efficient frontier in the (σ, μ)-space is highly sensitive to both the correlation coefficient ρ and whether short-selling is allowed. The figure below, illustrates the efficient frontier for three different values of the correlation coefficient: $\rho = -1, 0.5, 1$ where no short-selling is permitted. As ρ decreases, the efficient frontier becomes more "stretched" toward the expected return axis (μ-axis), reflecting increased diversification benefits.

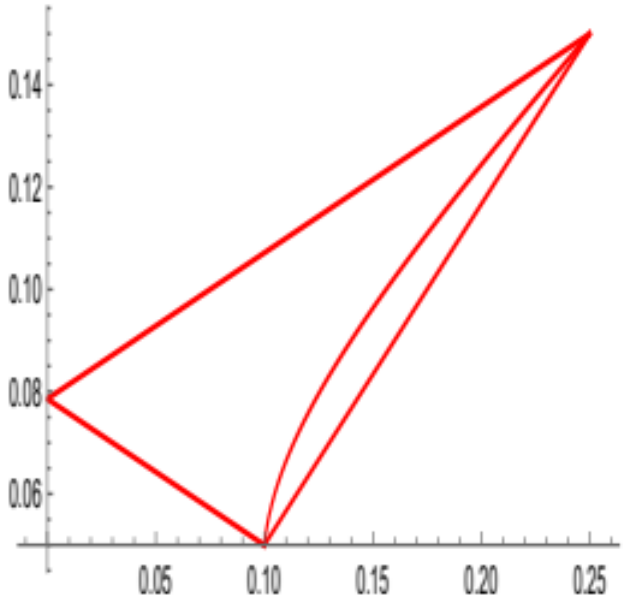


Figure 10: Frontiers for different correlations in the two-asset case.

Key Observations:

- **Short-selling** is excluded in the plot above. This leads to a more restricted set of attainable portfolios.
- **Lower correlation** between asset returns ($\rho \to -1$) increases the diversification benefit, thereby stretching the efficient frontier outward, especially in the direction of higher expected returns for a given level of risk.
- For a fixed level of portfolio risk, a lower correlation leads to a higher achievable return.
- The portfolio is said to be *more diversified* when $\rho \to -1$, as it allows for greater variance reduction via mixing.

Example: Capital Market Line and Efficient Frontier In the presence of a risk-free asset, the investor's opportunity set is no longer limited to the efficient frontier formed solely by risky assets. Instead, the optimal portfolio lies on the **Capital Market Line (CML)**, which represents combinations of the risk-free asset and the tangency portfolio of risky assets. The CML is a straight line from the risk-free rate (on the μ-axis) that is tangent to the efficient frontier.

In the figure below, we illustrate:

- the **efficient frontier** (blue curve), formed by all mean-variance efficient portfolios of risky assets,
- the **Capital Market Line** (red dashed line), which is tangent to the efficient frontier at the **tangency portfolio** (black dot),
- the **minimum variance portfolio** (green dot), which is the portfolio on the efficient frontier with the lowest variance.

The tangency portfolio is of special interest—it has the highest Sharpe ratio among all portfolios on the frontier and is the optimal risky portfolio for any investor who can combine it with the risk-free asset. Portfolios along the CML dominate those on the efficient frontier because they achieve higher expected returns for the same level of risk (or lower risk for the same expected return), thanks to the availability of the risk-free asset.

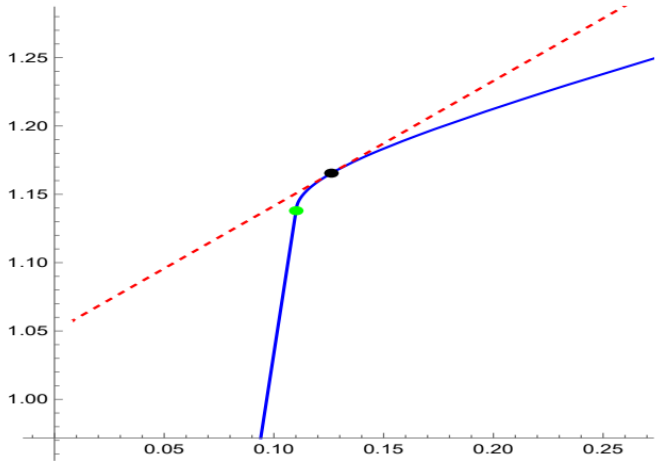


Figure 11: Efficient frontier and capital market line. The x-axis represents portfolio risk (standard deviation σ), and the y-axis represents expected return μ. The green point marks the minimum variance portfolio, and the black point marks the tangency portfolio.

One-Fund Theorem: This result leads to the so-called *One-Fund Theorem*, which states that all investors, regardless of their individual risk preferences, will hold some combination of:

- the risk-free asset, and
- a unique tangency portfolio of risky assets.

Sharpe Ratio

The **Sharpe Ratio** is a key performance metric in portfolio theory, measuring the excess return per unit of risk. It is defined as:

$$\text{Sharpe Ratio} = \frac{\mathbb{E}(r) - r_0}{\sqrt{\text{Var}(r)}},$$

where:

- $\mathbb{E}(r)$ is the expected return of the portfolio,
- r_0 is the return of the risk-free asset,
- $\text{Var}(r)$ is the portfolio return variance.

Assuming the total portfolio return is given by $r = \frac{V_1}{V_0}$, we can equivalently write:

$$\text{Sharpe Ratio} = \frac{\mathbb{E}(V_1) - r_0 V_0}{\sqrt{\text{Var}(V_1)}}.$$

Challenges in Parameter Estimation

In practical applications of portfolio optimization, the true values of the parameters $\boldsymbol{\mu}$ (expected returns) and $\boldsymbol{\Sigma}$ (covariance matrix) are generally unknown and must be estimated from historical data. This introduces a layer of uncertainty into the optimization process, as the resulting strategy is only as reliable as the estimates used.

Illustrative Example Consider the case where:

$$r_0 = V_0 = 1, \quad \boldsymbol{\mu} = \begin{pmatrix} \frac{41}{40} \\ \frac{43}{40} \end{pmatrix}, \quad \boldsymbol{\Sigma} = \begin{pmatrix} \frac{9}{100} & \frac{3}{40} \\ \frac{3}{40} & \frac{1}{4} \end{pmatrix}.$$

Using the parameters:

$$c = 0.5, \quad \sigma_0 = 0.3, \quad \mu_0 = 1.0451,$$

we find that all three optimization approaches—maximizing expected return under a risk constraint, minimizing variance under a return constraint, and solving the trade-off problem—yield approximately the same optimal strategy:

$$\mathbf{w} \approx \begin{pmatrix} 0.074 \\ 0.577 \end{pmatrix}.$$

Impact of Estimation Error While the analytical solution may appear stable and elegant, its real-world implementation is highly sensitive to estimation error. In particular:

- **Estimation error in expected returns $\boldsymbol{\mu}$** tends to have a large impact on the resulting portfolio weights, often leading to extreme positions when using plug-in estimates.
- **Covariance matrix estimation** is more stable, but still subject to sampling error, especially when the number of assets is large relative to the sample size.
- Optimization amplifies errors: small inaccuracies in $\boldsymbol{\mu}$ or $\boldsymbol{\Sigma}$ can lead to significantly different portfolios due to the nonlinear nature of the optimization.

Implications This sensitivity motivates the use of robust methods, such as:

- shrinkage estimators for $\boldsymbol{\Sigma}$,
- Bayesian or resampling-based approaches for $\boldsymbol{\mu}$,
- imposing constraints on portfolio weights to reduce overfitting to sample data.

In summary, while theoretical portfolio optimization assumes known parameters, real-world application must carefully consider the uncertainty in estimation and its consequences for risk and performance.

Drawbacks and Limitations of Mean-Variance Analysis

While mean-variance optimization is elegant and foundational in modern portfolio theory, it also suffers from several conceptual and practical limitations.

Limitations of Variance as a Risk Measure Variance (or its square root, standard deviation) is widely used to quantify risk as it measures the dispersion of returns around the mean—whether historical or expected. However, this measure has inherent drawbacks:

- Variance penalizes both upside and downside deviations equally. For a long-position investor, this is counterintuitive, as high positive returns are beneficial, not risky.
- The mean and variance together do not fully describe the return distribution, especially when the distribution is skewed or exhibits fat tails.
- The use of variance assumes that returns are approximately normally distributed. This assumption is only reasonable when the return distribution is symmetric and unimodal. In such cases, the mean and variance can be useful approximations; otherwise, they are insufficient.

Consequences and Practical Issues Due to these limitations, mean-variance analysis has not been universally adopted by practitioners. Key reasons include:

- **Variance is not a reliable risk measure** in all contexts, particularly because it does not distinguish between favorable and unfavorable deviations. This motivates the use of alternative risk measures such as:
  - downside deviation,
  - Value-at-Risk (VaR),
  - Conditional Value-at-Risk (CVaR),
  - semi-variance,
  - utility-based or loss-based risk functions.
- **High sensitivity to parameter estimation.** As previously discussed, portfolio weights derived from mean-variance optimization can vary dramatically with small changes in the estimates of expected returns or covariances. This instability often leads to extreme and unrealistic allocations.

  A common remedy is the use of:
  - robust optimization techniques that account for estimation uncertainty,
  - shrinkage estimators or Bayesian priors for more stable covariance and return estimates,
  - constraints on weights to reduce overfitting to sample data.

These observations motivate the development of extended portfolio models that incorporate better risk measures, account for non-normality, and improve robustness in the presence of parameter uncertainty.

Mean-Value-at-Risk (Mean-VaR) Analysis

In the classical Markowitz framework, portfolio efficiency is evaluated in terms of expected return and variance (or standard deviation). However, variance penalizes both upward and downward deviations equally, which motivates the use of alternative risk measures that better capture downside risk.

One popular alternative is the **Value-at-Risk (VaR)**, which provides a quantile-based measure of potential loss. That is, for a confidence level $\rho \in (0, 1)$, the VaR at level ρ is the maximum loss not exceeded with probability ρ.

VaR in Normally and t-Distributed Returns When returns are assumed to follow a **multivariate normal distribution**, the Value-at-Risk of the portfolio return r_p is given by:

$$\mathrm{VaR}_\rho(r_p) = \Phi^{-1}(1-\rho) \cdot \sqrt{\mathrm{Var}(r)} - \mathbb{E}(r),$$

where Φ^{-1} is the quantile function of the standard normal distribution.

If the returns instead follow a **Student's t-distribution** with ν degrees of freedom, the expression becomes:

$$\mathrm{VaR}_\rho(r_p) = t_\nu^{-1}(1-\rho) \cdot \left(\frac{\nu-2}{\nu}\right)^{1/2} \cdot \sqrt{\mathrm{Var}(r)} - \mathbb{E}(r),$$

where t_ν^{-1} is the quantile function of the t-distribution. In either case, we define:

$$\mathrm{VaR}_\rho(r_p) = z^* \sigma_p - \mu_p,$$

where z^* depends on the assumed distribution and confidence level ρ.

Linking VaR to the Efficient Frontier Using the relationship $\sigma_p = \frac{\mathrm{VaR}_\rho(r_p)+\mu_p}{z^*}$, we can plug this into the efficient frontier equation. The efficient frontier expressed in the (VaR_ρ, μ)-plane becomes:

$$\frac{\mathrm{VaR}_\rho(r_p) + \mu_p}{z^*} = \sqrt{\frac{c\mu_p^2 - 2a\mu_p + b}{d}},$$

where:

$$\begin{aligned} a &= \mathbf{1}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}, \\ b &= \boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}, \\ c &= \mathbf{1}^\top \boldsymbol{\Sigma}^{-1} \mathbf{1}, \\ d &= bc - a^2. \end{aligned}$$

This representation shows how we can derive the efficient frontier under VaR-based risk assessment. While VaR is not a convex function in general (which complicates optimization), under normal or t-distributed returns, it becomes tractable and provides a useful downside-focused risk measure.

Remarks

- VaR in this context is defined in **monetary units** (rather than percentages), since it reflects potential loss on the value of the portfolio.
- The expressions mirror the mean-variance frontier structure, except that variance is replaced with VaR.
- This allows for more intuitive risk communication (e.g., "there is a 95% chance we won't lose more than X").

Efficient Frontiers in Mean-VaR Analysis

The relationship between efficient frontiers in the mean-variance and mean-VaR frameworks is both theoretically insightful and practically relevant.

Inclusion Relationship Between Frontiers It can be shown that:

- Every portfolio that is efficient under the mean-VaR criterion is also efficient under the mean-variance criterion.
- However, the reverse is not true: not all mean-variance efficient portfolios are mean-VaR efficient.

Let $\mathbf{w}^*$ denote a mean-variance efficient portfolio. If this portfolio lies between two mean-VaR efficient portfolios $r_{m_{VaR_1}}$ and $r_{m_{VaR_2}}$ such that:

$$\mathbb{E}(r_{m_{VaR_2}}) < \mathbb{E}(\mathbf{w}^*) < \mathbb{E}(r_{m_{VaR_1}}),$$

then $\mathbf{w}^*$ is not efficient under the mean-VaR framework.

Consequently, the set of mean-VaR efficient portfolios is a *proper subset* of the set of mean-variance efficient portfolios:

$$\text{Eff}_{\text{VaR}} \subset \text{Eff}_{\text{MV}}.$$

This result arises because mean-VaR analysis imposes stricter constraints due to its focus on downside risk (quantile-based), which leads to the exclusion of some portfolios considered efficient by mean-variance analysis.

Behavior for Extreme Confidence Levels The nature of the mean-VaR efficient frontier also depends on the confidence level ρ. In particular, the frontier only exists for relatively low values of $p = 1 - \rho$, i.e., high confidence levels when measuring loss.

As $p \to 0$ (confidence $\rho \to 1$), we observe the following behavior:

$$z_p^* = \Phi^{-1}(1-p) \to \infty, \quad \frac{\text{VaR}_p}{z_p^*} = \sigma_p + \frac{\mu_p}{z_p^*} \to \sigma_p.$$

This means that the VaR increasingly resembles the standard deviation for very small values of p, and the **minimum VaR portfolio converges to the minimum variance portfolio**:

$$\lim_{p \to 0} \text{VaR}_{\min} = \text{Variance}_{\min}.$$

This limiting result demonstrates the consistency of the two approaches under extreme risk-aversion scenarios and helps validate the use of variance as an approximation in such contexts.

12 Problem by topic:

12.1 Finance Concepts:

- Exercise 1: 1 (Returns, inflation and CPI)

12.2 Basics of Statistical Inference

- Preliminary Exercise: 1 (Moments)(Done)
- Preliminary Exercise: 2 (Method of moments and MLE Estimation) (Done)
- Preliminary Exercise: 3 (MSE and Consistency) (Done)
- Preliminary Exercise: 4 (MSE) (Done)
- Exam 2023: (Defining formulas) 1ac (Done)

12.3 Regression Techniques

- Exercise 1: 2 (Derive beta estimators) (Done)
- Exercise 1: 3 (PRF vs SRF) (Done)
- Exercise 1: 4 (assumptions error terms vs sampled) (Done)
- Exercise 1: 5 (conditional expectation linear regression) (Done)
- Exercise 1: 7 (type 1 and 2 error) (Done)
- Exercise 1: 8 (hypothesis testing coefficients) (Done)

- Exercise 1: 9 (Setup regression and re-write with logs) (Done)
- Exercise 2: 1 (t-test or F-test) (Done)
- Exercise 2: 2 (RRSS vs URSS) (Done)
- Exercise 2: 3 (find restricted regression, carry out F-test) (Done)
- Exercise 2: 4 (t-ratios and parameter significance) (Done)
- Exercise 2: 5-9 (Programming) (Done)
- Exercise 2: 10 (R-squared) (Done)
- Voluntary Exercise 1: 1 (Why little attention to assumption 1 CLRM) (Done)
- Voluntary Exercise 1: 2 (Assumption 2 CLRM discussion) (Done)
- Voluntary Exercise 1: 3 (Heteroscedasticity tests) (Done)
- Voluntary Exercise 1: 4 (DW test and others for autocorrealtion) (Done)
- Voluntary Exercise 1: 5 (Cochrane-Orcutt method and assumptions) (Done)
- Voluntary Exercise 1: 6 (Stochasticity of explanatory variables) (Done)
- Voluntary Exercise 1: 7 (Exogeneity assumption, fixes) (Done)
- Voluntary Exercise 1: 8 (BJ test understanding) (Done)
- Voluntary Exercise 1: 9 (parameter stability test, why) (Done)
- Exam 2024: 1.1 (Set up regression model, and interpret outputs from regression) (Done)
- Exam 2024: 1.2 (IV understanding and estimation) (Done)
- Exam 2023: 3abcd
- Exam 2020: 3.1 (Multicollinearinty) (Done)
- Exam 2020: 3.2 (R-squared and unit understanding) (Done)
- Exam 2020: 4.1 (Set up regression, and how interaction term changes coefficient interpretations) (Done)
- Exam 2019: 4 (Assumptions of CLRM, paramter interpretation and multiple linear regression model) (Done)

12.4 Time Series Analysis

- Exercise 3: 1 (Differences AR and MA) (Done)
- Exercise 3: 2 (AR and MA acf, pacf and memory of shocks) (Done)
- Exercise 3: 3 (Determine process from sample acf pacf, Ljung-Box test) (Done)
- Exercise 3: 4 (Check model for stationarity) (Done)
- Exercise 3: 5 (Select model based on information criterion) (Done)
- Exercise 3: 6 (Forecasting ARMA, and MSE for eval) (Done)
- Exercise 3: 7 (Programming) (Done)
- Exercise 3: 8 (How to make time series stationary, importance of stationarity for inference and define time series terms) (Done)
- Exercise 4: 1-8 (Read through prints and understand) (Done)
- Exam 2024: 2.1 (Stylized facts about financial asset prices and returns, and why ARMA cannot model them) (Done)
- Exam 2024: 2.2 (ARIMA specifications) (Done)
- Exam 2024: 2.3 (AIC and BIC) (Done)
- Exam 2023: 1d (acf, pacf and q-test)
- Exam 2023: 4abcdef (Why ARMA models require stationarity, weakly stationary conditions, ARMA vs ARIMA, random walk and non-stationarity, significance of coefficients and predictive power, IC's) (Done)
- Exam 2019: 3 (Same as exercise 3: 8) (Done)

12.5 Volatility models

- Exercise 5: 1-8 (Read through and understand prints: Important problem) (Done)
- Exercise 6: 1 (Multivariate GARCH) (Done)
- Exam 2024: 3.1 (Define ARMA-ARCH model mathematically and find kurtosis of arch(1)) (Done)
- Exam 2024: 3.2 (Specify model based on R print and comment on result: important problem) (Done)
- Exam 2023: 4abcd (Describe Garch models)

12.6 Multivariate models

- Exercise 6: 2.1 (Define VAR and its general equation) (Done)
- Exercise 6: 2.2 (Advantages and limitations of VARs) (Done)
- Exercise 6: 2.3 (Optimal lag length VAR) (Done)
- Exercise 6: 2.4 (Granger Causality) (Done)
- Exercise 6: 2.5 (Impulse response functions) (Done)
- Exam 2020: 3.3 (Optimal lag length testing) (Done)
- Exam 2020: 3.4 (Cointegration of VAR) (Done)
- Exam 2020: 3.5 (Granger causality checks) (Done)

12.7 Panel Data Analysis

- Exercise 7: 1 (Check FE and RE significane and state R-squared meanings) (Done)
- Exercise 7: 2 (Write equations, assumptions and estimation methods for FE and RE and highlight main difference) (Done)
- Exercise 7: 3 (Based on output: preferred model) (Done)
- Exercise 7: 4 (what does σ_u, σ_e and ρ stand for in regression output. What does F-test represent in FE model output) (Done)
- Exercise 7: 5 (Why should we use cluster-robust standard errors) (Done)
- Exercise 7: 6 (Huasman test for best model) (Done)
- Exam 2024: 3.3 (Estimators for FE and RE) (Done)

12.8 Financial Risk Management

- Exercise 8: 1 (Portfolio construction, rho-calculation and portoflio composition) (Done)
- Exercise 8: 2 (Subadditivity desirable) (Done)
- Exercise 9: 1 (Caluclate VaR for a distribution and find sigma for normal using VaR value. Discuss limitation of VaR, and calculate ES.) (Done)
- Exercise 9: 2 (Given a portfolio, prove that VaR is not subadditive for it) (Done)
- Exercise 9: 3 (Empirical confidence interval) (Done)

- Extra Exercise: 1 (Deduce portfolio value, minimize mean-variance risk, find VaR, given weights, find ES from VaR, discuss results) (Done)
- Extra Exercise: 2 (t-distribution of VaR converges to normal distrbution with increasing degrees of freedom) (Done)
- Extra Exercise: 3 (Employ formulas for VaR and ES of t-dsitrbution) (Done)
- Extra Exercise: 4 (Empirical VaR and ES= (Done)
- Extra Exercise: 6 (Disadvantage of mean-VaR analysis (convexity)) (Done)
- Exercise 10: 1 (Comment on correlation coefficient with regards to efficient frontier and portfolio returns, calculate minimum variance portfolio, calculate capital market line)
- Exercise 10: 2 (Limitations on mean-variance approach) (Done)
- Exam 2024: 4.1 (Model portfolio using annual returns, find VaR, prove its subadditive, and discuss subadditivity in context of VaR and finaincial importance) (Done)
- Exam 2024: 4.2 (Analyzing efficieny of portfolio in mean-VaR analysis and mean-variance analysis) (Done)
- Exam 2023: 1e (Empirical var and ES calculations) (Done)