Contents

Part I

Quantitative Analysis

Chapter 1

Things to understand

- how variance of the mean and variance is derived
- how to generate correlated random numbers
- why R^2 is the squared correlation coefficient between Y and X
- ways to handle multi-collinearity
- Bonferroni method
- fit metrics:
  - MSE - mean squared error
  - R^2
  - SER - standard error of the regression
  - SSR - sum of squared residuals
  - TSS - total sum of squares
  - ESS - explained sum of squares
- Durbin-Watson statistics

- the difference between auto-correlation and partial auto-correlation
- Box-Pierce Q-statistic
- Ljung-Box Q-statistic
- Dickey-Fuller test
- unit root test
- bias-variance decomposition

Chapter 2

Probability

2.1 Conditional Probability

$$P(A|B) \quad = \quad \frac{P(A \cap B)}{P(B)} \tag{2.1}$$

2.2 Bayes Theorem

Writing

$$P(A \cap B) \quad = \quad P(A|B)P(B) = P(B|A)P(A) \tag{2.2}$$

we get:

$$P(A|B) \quad = \quad \frac{P(B|A)P(A)}{P(B)} \tag{2.3}$$

2.3 Law of Total Probability

If $\{B_n : n = 1, 2, 3, ...\}$ is a set of **pairwise disjoint event whose union is the entire sample space** and each event is measurable, then for any event A of the same probability space:

$$P(A) = \sum_n Pr(A|B_n)P(B) \tag{2.4}$$

2.4 Bayesian vs Frequentists

Frequentist approach counts positive/negative outcomes and based on this derives probability.

Bayesian approach starts with a prior belief about the probability. In most cases the prior is either subjective or based on frequentist analysis.

Situations in which there is very little data, or in which the signal-to-noise ratio is very low, often require Bayesian analysis. When we have lots of data,the conclusions of frequentist analysis and Bayesian analysis are often very similar, and the frequentist results are often easier to calculate.

Chapter 3

Distribution Descriptive Statistics

3.1 Expectation

Expectation is a linear operator:

$$\mathbb{E}(X+Y) = \mathbb{E}(X)+\mathbb{E}(Y) \tag{3.1}$$

$$\mathbb{E}(c\cdot X) = c\cdot\mathbb{E}(X) \tag{3.2}$$

3.2 Moment

The n-th moment of a real-valued continuous function $f(x)$ of a real variable about a value c is:

$$\mu_n = \int_{-\infty}^{\infty}\Big(x-c\Big)^n f(x)dx \tag{3.3}$$

If $c = 0$, the moments are called **raw moments**.

If $c = \mu$, where μ is mean of the distribution, the moments are called **central moments**.

The moment

$$\mu_n = \int_{-\infty}^{\infty} \left(\frac{x - c}{\sigma} \right)^n f(x) dx \tag{3.4}$$

is called **standardised moment**.

If we have a data set and we want to standardise it, we first compute the sample mean ans the standard deviation. Then, for each data point, we subtract the mean and divide by the standard deviation.

The inverse transformation can also be very useful when it comes to creating simulations. Simulations often begin with standardised variables, which need to be transformed into variables with a specific mean and standard deviation. In this case, we simply take the output from the standardised variable, multiply by the desired standard deviation, and then add the desired mean. The order is important. Adding a constant to a random variable will not change the standard deviation, but multiplying a non-mean-zero variable by a constant will change the mean.

3.3 Mean

Mean is the first raw moment:

$$\mu = \mathbb{E}(X) \tag{3.5}$$

For a discrete variable it is:

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{3.6}$$

3.3.1 Properties

Mean is a linear function.

3.4 Variance

Variance is the second central moment:

$$\sigma^2 = \mathbb{E}(X - \bar{X})^2 \tag{3.7}$$

$$= \mathbb{E}\Big(X^2\Big) - \Big(\mathbb{E}(X)\Big)^2 \tag{3.8}$$

$$= \mathbb{E}\Big(X^2\Big) - \bar{X}^2 \tag{3.9}$$

Sample unbiased variance for a discrete variable is:

$$\hat{\sigma}^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{X})^2 \tag{3.10}$$

Sample maximum likelihood variance for a discrete variable is:

$$\hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{X})^2 \tag{3.11}$$

3.4.1 Properties

- Multiplying a random variable by a constant c will increase the variance by c^2:

$$\sigma^2(c \cdot X) \quad = \quad c^2\sigma^2(X) \tag{3.12}$$

- Adding a constant c to a random variable will not change the standard deviation or variance of the distribution:

$$\sigma(X + c) \quad = \quad \sigma(X) \tag{3.13}$$

3.5 Skewness

Skewness is the third standardised central moment:

$$\gamma_1 \quad = \quad s = \text{Skewness} = \mathbb{E}\left[\left(\frac{X - \mu}{\sigma}\right)^3\right] \tag{3.14}$$

Unbiased sample skewness is defined as:

$$\hat{s} \quad = \quad \frac{N}{(N-1)(N-2)} \sum_{i=1}^{N} \left(\frac{x_i - \mu}{\sigma}\right)^3 \tag{3.15}$$

3.5.1 Properties

- Multiplying a random variable by a constant will not change the value of the skewness of the distribution:

$$s(c \cdot X) \quad = \quad s(X) \tag{3.16}$$

- the formula in terms of $\mathbb{E}\Big(X^3\Big)$ and μ is not as simple as that for the variance:

$$\mathbb{E}\left[\Big(X-\mu\Big)^3\right] \quad = \quad \mathbb{E}\Big(X^3\Big) - 3\mu\sigma^2 - \mu^3 \tag{3.17}$$

3.6 Kurtosis

Kurtosis is the fourth standardised central moment.

$$K \quad = \quad \mathbb{E}\left[\left(\frac{X-\mu}{\sigma}\right)^4\right] \tag{3.18}$$

Unbiased sample kurtosis is defined as:

$$\hat{K} \quad = \quad \frac{N(N+1)}{(N-1)(N-2)(N-3)} \sum_{i=1}^{N} \left(\frac{x_i-\mu}{\sigma}\right)^4 \tag{3.19}$$

3.6.1 Kurtosis Excess

Kurtosis excess is the difference between kurtosis of the distribution and the kurtosis of the normal distribution:

$$K_{excess} = K - K_{Normal} = \mathbb{E}\left[\left(\frac{X-\mu}{\sigma}\right)^4\right] - 3 \tag{3.20}$$

Sample unbiased kurtosis excess is defined as:

$$\hat{K}_{excess} = \hat{K} - 3\frac{(N-1)^2}{(N-2)(N-3)} \tag{3.21}$$

3.6.2 Properties

- as with skewness, multiplying a random variable by a constant c will not change the kurtosis

$$K(c \cdot X) = K(X) \tag{3.22}$$

- if the distribution of returns of two assets have the same mean, variance and skewness but different kurtosis, then the distribution with the higher kurtosis will tend to have more extreme points and be considered more risky

3.7 Zero Scores

The z-score is calculated as

$$z = \frac{X - \mu}{\sigma} \tag{3.23}$$

It therefore gives the location of raw scores above and below the mean in units of the standard deviation of the distribution from the mean.

3.8 Continuity Correction

A continuity correction is an adjustment that is made when a discrete distribution is approximated by a continuous distribution.
If X is the original discrete distribution that is approximated by Y continuous distribution, then:

$$P\big(X \leq Y\big) \approx P\big(Y \leq x + \frac{1}{2}\big) \tag{3.24}$$

Chapter 4

Uniform Distribution

Continuous uniform or rectangular distribution is a family of symmetric probability distributions, such that for each member of the family, all intervals of the same length on the distribution support are equally probable.
The support is defined by two parameters, a and b, which are its minimum and maximum values.

4.1 Probability Density Function

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{for } a \leq x \leq b \\ 0, & \text{for } x < a \text{ or } x > b \end{cases} \tag{4.1}$$

In terms of mean μ and variance σ the p.d.f. may be written as:

$$f(x) = \begin{cases} \frac{1}{2\sigma\sqrt{3}} & \text{for } -\sigma\sqrt{3} \leq x - \mu \leq \sigma\sqrt{3} \\ 0 & \text{otherwise} \end{cases} \tag{4.2}$$

These can be written in terms of the Heaviside step function $H(x)$ as:

$$f(x) = \frac{H(x-a) - H(x-b)}{b-a} \tag{4.3}$$

4.2 Cumulative Distribution Function

$$F(x) = \begin{cases} 0 & \text{for } x < a \\ \frac{x-a}{x-b} & \text{for } a \leq x \leq b \\ 1 & \text{for } x > b \end{cases} \tag{4.4}$$

Its inverse is:

$$F^{-1}(p) = a + p(b-a) \text{ for } 0 < p < 1 \tag{4.5}$$

In mean and variance notation, the cumulative distribution function is:

$$F(x) = \begin{cases} 0 & \text{for } x < \mu - \sigma\sqrt{3} \\ \frac{1}{2}\left(\frac{x-\mu}{\sigma\sqrt{3}} + 1\right) & \text{for } \mu - \sigma\sqrt{3} \leq x \leq \mu + \sigma\sqrt{3} \\ 1 & \text{for } x > \mu + \sigma\sqrt{3} \end{cases} \tag{4.6}$$

And its inverse is:

$$F - 1(p) = \sigma\sqrt{3}(2p - a) + \mu \text{ for } 0 \leq p \leq 1 \tag{4.7}$$

In the terms of the Heaviside step function $H(x)$:

$$F(x) = \frac{(x-a)H(x-a) - (x-b)H(x-b)}{b-a} \tag{4.8}$$

4.3 Generating Functions

Moment Generating Function The moment generating function is:

$$\begin{aligned} M_x &= \mathbb{E}\Big(e^{tx}\Big) = \int_a^b \frac{e^{tx}}{b-a}dx = \frac{e^{tx}}{t(b-a)}\Big|_a^b \\ &= \begin{cases} \frac{e^{tb}-e^{ta}}{t(b-a)} & \text{for } t \neq 0 \\ 1 & \text{for } t = 0 \end{cases} \end{aligned} \tag{4.9}$$

4.4 Moments

The moment-generating function is not differentiable at zero, but the moments can be calculated by differentiating and then taking $\lim_{t\to 0}$. However, it is easier to calculate the moments from the definition directly.

Row Moments

$$\begin{aligned} \mu_n &= \int_{-\infty}^{\infty} \frac{H(x-a) - H(x-b)}{b-a} x^n dx \\ &= \int_a^b \frac{x^n}{b-a} dx \\ &= \frac{b^{n+1} - a^{n+1}}{(n+1)(b-a)} \end{aligned} \tag{4.10}$$

The first few are:

$$\mu_1 = \frac{1}{2}(a+b) \tag{4.11}$$

$$\mu_2 = \frac{1}{3}(a^2+ab+b^2) \tag{4.12}$$

$$\mu_3 = \frac{1}{4}(a+b)(a^2+b2) \tag{4.13}$$

$$\mu_4 = \frac{1}{5}(a^4+a^3b+a^2b^2+ab^3+b^4) \tag{4.14}$$

Central Moments The central moments are given analytically by:

$$\begin{aligned}\mu_n &= \int_{-\infty}^{\infty} \frac{H(x-a)-H(x-b)}{b-a}\left[x-\frac{1}{2}(a+b)\right]^n dx \\ &= \int_a^b \frac{\left[x-\frac{1}{2}(a+b)\right]^n}{b-a} ds \\ &= \frac{\Big(a-b\Big)^n+\Big(b-a\Big)^n}{2^{n+1}\Big(n+1\Big)}\end{aligned} \tag{4.15}$$

The first few are:

$$\mu_1 = 0 \tag{4.16}$$

$$\mu_2 = \frac{1}{12}(b-a)^2 \tag{4.17}$$

$$\mu_3 = 0 \tag{4.18}$$

$$\mu_4 = \frac{1}{80}(b-a)^4 \tag{4.19}$$

4.5 Mean, Variance, Skew, Kurtosis Excess

Mean (the first raw moment) is:

$$\mu = \frac{1}{2}(a+b) \tag{4.20}$$

Variance (the second central moment) is:

$$\sigma^2 = \frac{1}{12}(b-a)^2 \tag{4.21}$$

Skew (third central moment normalised by standard deviation σ) is:

$$\gamma_1 = 0 \tag{4.22}$$

Kurtosis excess (the fourth central moment normalised by standard) is:

$$\gamma_2 = -\frac{6}{5} \tag{4.23}$$

Chapter 5

Normal Distribution

Chapter 6

Lognormal Distribution

Chapter 7

Bernoulli Distribution

The Bernoulli distribution is a discrete distribution having two possible outcomes labelled by $n = 0$ and $n = 1$. In which

- $n = 1$ ("success") occurs with probability p
- and $n = 0$ ("failure") occurs with probability $q = 1 - p$

where $0 < p < 1$.

7.1 Probability Density Function

It therefore has probability density function

$$P(n) = \begin{cases} 1-p, & \text{for } n = 0 \\ p, & \text{for } n = 1 \end{cases} \tag{7.1}$$

7.2 Cumulative Distribution Function

$$F(n) = \begin{cases} 1-p, & \text{for } n = 0 \\ 1, & \text{for } n = 1 \end{cases} \tag{7.2}$$

7.3 Mean

Let X be a discrete random variable with the Bernoulli distribution with parameter p.

Then the expectation of X is given by:

$$\mathbb{E}\big(X\big) = p \tag{7.3}$$

Proof From the definition of expectation:

$$\mathbb{E}\big(X\big) = \sum_{x \in \Omega_X} x Pr\big(X = x\big) \tag{7.4}$$

By definition of Bernoulli distribution:

$$\mathbb{E}\big(X\big) = 1xp + 0x(1-p) = p \tag{7.5}$$

7.4 Variance

The variance of X is given by:

$$\sigma^2\big(X\big) \quad = \quad p\big(1-p\big) \tag{7.6}$$

Proof From the definition of variance:

$$\sigma^2\big(X\big) \quad = \quad \mathbb{E}\bigg(\Big(X-\mathbb{E}\big(X\big)\Big)^2\bigg)$$

From the expectation of Bernoulli distribution, we have $\mathbb{E}\big(X\big)=p$

$$\begin{aligned}
\sigma^2\big(X\big) &= \mathbb{E}\bigg(\Big(X-\mathbb{E}\big(X\big)\Big)^2\bigg) \\
&= \big(1-p\big)^2 p + \big(0-p\big)^2\big(1-p\big) \\
&= p - 2p^2 + p^3 + p^2 - p^3 \\
&= p(1-p) \qquad (7.7)
\end{aligned}$$

The derivation from "Variance equals expectation of square minus square of expectation" is a bit simpler.

7.5 Moments - Summary

$$\mu = p \tag{7.8}$$

$$\sigma^2 = p(1-p) \tag{7.9}$$

$$\gamma_1 = \frac{1-2p}{\sqrt{p(1-p)}} \tag{7.10}$$

$$\gamma_2 = \frac{6p^2-6p+1}{p(1-p)} \tag{7.11}$$

Chapter 8

Binomial Distribution

The binomial distribution with parameters n and p is the **discrete probability distribution** $P_p(n|N)$ of obtaining exactly n successes out of N **Bernoulli trials** (where the result of each **Bernoulli trials** is true with probability p and false with probability $q = 1 - p$).

The binomial distribution is therefore given by

$$P_p(n|N) = \binom{N}{n} p^n \cdot q^{N-n} = \frac{N!}{n!(N-n)!} p^n \cdot q^{N-n} \tag{8.1}$$

where $\binom{N}{n}$ is a **binomial coefficient**.

8.1 Properties

- the distribution is symmetrical when $p = 0.5$ or when n is large
- the distribution could be thought as a collection of Bernoulli random variables

8.2 Mean

Let X be a discrete random variable with the binomial distribution with parameters n and p.

Then the expectation of X is given by:

$$\mathbb{E}(X) = np \tag{8.2}$$

Proof From the definition of expectation

$$\mathbb{E}(X) = \sum_{x \in \Omega_X} x Pr(X = x) \tag{8.3}$$

From **Bernoulli Process as Binomial Distribution**, we see that X is a sum of **discrete random variables** Y_i that model the **Bernoulli Distribution**:

$$X = \sum_{i=1}^{n} Y_i \tag{8.4}$$

Each of the **Bernoulli trials** is independent of each other, by definition of a **Bernoulli process**. It follows that:

$$\begin{aligned}
\mathbb{E}(X) &= \mathbb{E}\left(\sum_{i=1}^{n} Y_i\right) \\
&= \sum_{i=1}^{n} \mathbb{E}\left(Y_i\right) && \text{sum of expecations of independent trials} \\
&= \sum_{i=1}^{n} p && \text{expectation of Bernoulli distribution} \\
&= np && \text{sum of identical terms}
\end{aligned} \tag{8.5}$$

8.3 Variance

Let X be a discrete random variable with the binomial distribution with parameters n and p.

Then the variance of X is given by:

$$\sigma^2\big(X\big) \quad = \quad np(1-p) \tag{8.6}$$

Proof From Binomial distribution as a sum of independent Bernoulli variables.

$$\begin{aligned} var\big(X\big) &= var(\sum_{i=1}^{n} Y_i) \\ &= \sum_{i=1}^{n} var(Y_i) && \text{from sum of variances of independent trials} \\ &= \sum_{i=1}^{n} p(1-p) && \text{variance of Bernoulli distribution is } p(1-p) \\ &= np(1-p) && (8.7) \end{aligned}$$

Chapter 9

Poisson Distribution

Poisson distribution is a discrete probability distribution that defines the probability of a given number of events occurring in a fixed interval of time or space if these events occur with a known constant rate and independently of the time since the last event. The Poisson distribution can also be used for the number of events is other specified intervals such as distance, area or volume.

The average number of events in an interval is designed as λ. Lambda is the event rate, also called the rate parameter. The probability of observing k events in an interval is given by the equation:

$$P(\text{k events in interval}) = \frac{\lambda^k}{k!}e^{-\lambda} \tag{9.1}$$

This equation can be adapted if, instead of the average number of events λ, we are given a time rate r for the events to happen. Then $\lambda = rt$ (with r in units of $1/time$), and

$$P(\text{k events in interval } t) = e^{-rt}\frac{(rt)^k}{k!} \tag{9.2}$$

The Poisson distribution could be obtained as the limit from the binomial distribution. The probability of obtaining exactly n successes in N trials is:

$$P_p\big(n\big|N\big) \quad = \quad \frac{N!}{n!(N-n)!}p^n(1-p)^{N-n} \tag{9.3}$$

Viewing the distribution as a function of the expected number of successes:

$$\lambda \quad = \quad Np \tag{9.4}$$

instead of the sample size N for fixed p, the equation above then becomes:

$$P_\lambda\big(n\big|N\big) \quad = \quad \frac{N!}{n!(N-n)!}\frac{\lambda}{N}^n\Big(1-\frac{\lambda}{N}\Big)^{N-n} \tag{9.5}$$

Letting the sample size N become large, the distribution then approaches:

$$\begin{aligned} P_\lambda(n) &= \lim_{N\to\infty} P_p \\ &= \lim_{N\to\infty} \frac{N(N-1)...(N-n+1)}{n!}\frac{\lambda^n}{N^n}\left(1-\frac{\lambda}{N}\right)^N\left(1-\frac{\lambda}{N}\right)^{-n} \\ &= \lim_{N\to\infty} \frac{N(N-1)...(N-n+1)}{N^n}\frac{\lambda^n}{n!}\left(1-\frac{\lambda}{N}\right)^N\left(1-\frac{\lambda}{N}\right)^{-n} \\ &= 1\cdot\frac{\lambda^n}{n!}\cdot e^{-\lambda}\cdot 1 \\ &= \frac{\lambda^n}{n!}e^{-\lambda} \end{aligned} \tag{9.6}$$

Note that the sample size N has completely dropped out of the probability function, which has the same functional form for all values of λ.

In this case, n is the number of events that occur in an interval, and λ is the expected number of events in the interval.

If the rate at which events occur over time is constant, and the probability of any one event occurring is independent of all other events, then we say that the

events follow a Possion process, where

$$P[X = n] \quad = \quad \frac{(\lambda t)^n}{n!} e^{-\lambda t} \tag{9.7}$$

where t is the time elapsed. In other words, the expected number of events before time t is equal to λt.

9.1 Properties

- the Poisson distribution is normalised so that the sum of probabilities equals 1 since:

$$\sum_{n=0}^{\infty} P_\lambda(n) = e^{-\lambda} \sum_{n=0}^{\infty} \frac{\lambda^n}{n!} = e^{-\lambda} e^{\lambda} = 1 \tag{9.8}$$

9.2 Mean

The expectation of X is given by:

$$\mathbb{E}\big(X\big) \quad = \quad \lambda \tag{9.9}$$

Proof: From the definition of expectation

$$\mathbb{E}\big(X\big) \quad = \quad \sum_{x \in \Omega_X} x \cdot Pr\big(X = x\big) \tag{9.10}$$

By definition of Poisson distribution

$$\begin{aligned}
\mathbb{E}(X) &= \sum_{k\geq 0} k\frac{1}{k!}\lambda^k e^{-\lambda} \\
&= \lambda e^{-\lambda}\sum_{k\geq 1}\frac{1}{(k-1)!}\lambda^{k-1} && \text{as the } k=0 \text{ term vanishes} \\
&= \lambda e^{-\lambda}\sum_{j\geq 0}\frac{1}{j!}\lambda^{j} && \text{putting } j=k-1 \\
&= \lambda e^{-\lambda}e^{\lambda} && \text{Taylor series expansion for exponential function} \\
&= \lambda
\end{aligned}$$

9.3 Variance

$$\sigma^2 = \lambda \tag{9.11}$$

Proof: From the definition of **Variance as Expectation of Square Minus Square of Expectation**:

$$var(X) = \mathbb{E}(X^2) - \Big(\mathbb{E}(X)\Big)^2 \tag{9.12}$$

From the **expectation of function of discrete random variable**:

$$\mathbb{E}(X^2) = \sum_{x\in\Omega_X} x^2 Pr(X=x) \tag{9.13}$$

So:

$$\begin{aligned}
\mathbb{E}\big(X^2\big) &= \sum_{k\geq 0} k^2 \frac{1}{k}\lambda^k e^{-\lambda} && \text{definition of Poisson Distribution} \\
&= \lambda e^{-\lambda} \sum_{k\geq 1} k \frac{1}{\big(k-1\big)!}\lambda^{k-1} && \text{note change of limit: term is zero when k = 0} \\
&= \lambda e^{-\lambda}\left(\sum_{k\geq 1}(k-1)\frac{1}{\big(k-1\big)!}\lambda^{k-1} + \sum_{k\geq 1}\frac{1}{\big(k-1\big)!}\lambda^{k-1}\right) \\
&= \lambda e^{-\lambda}\left(\lambda\sum_{k\geq 2}\frac{1}{(k-2)!}\lambda^{k-2} + \sum_{k\geq 1}\frac{1}{\big(k-1\big)!}\lambda^{k-1}\right) \\
&= \lambda e^{-\lambda}\left(\lambda\sum_{i\geq 0}\frac{1}{i!}\lambda^{i} + \sum_{j\geq 0}\frac{1}{j!}\lambda^{j}\right) && \text{putting } i = k-2,\ j = k-1 \\
&= \lambda e^{-\lambda}\Big(\lambda e^{\lambda} + e^{\lambda}\Big) && \text{Taylor series expansion for exponential function} \\
&= \lambda\big(\lambda+1\big) = \lambda^2 + \lambda && (9.14)
\end{aligned}$$

Then:

$$\begin{aligned}
var(X) &= \mathbb{E}\Big(X^2\Big) - \Big(\mathbb{E}\big(X\big)\Big)^2 \\
&= \lambda^2 + \lambda - \lambda^2 = \lambda && (9.15)
\end{aligned}$$

9.4 Moments - Summary

$$\mu = \lambda \tag{9.16}$$

$$\sigma^2 = \lambda \tag{9.17}$$

$$\gamma_1 = \sqrt{\lambda} \tag{9.18}$$

$$\gamma_2 = \frac{1}{\lambda} \tag{9.19}$$

Chapter 10

Chi-squared Distribution

Chi-squared distribution (χ^2-distribution) with k degrees of freedom is the distribution of a **sum of the squares of k independent standard normal** variables.

When it is being distinguished from the more general *non-central chi-squared distribution*, this distribution is sometimes called the **central chi-squared distribution**.

10.1 Use

- common **chi-squared tests for goodness of fit** of an observed distribution to a theoretical one
- the **independence** of two criteria of classification of qualitative data
- in **confidence interval estimation for a population standard deviation of a normal distribution** from a sample standard deviation
- **Friedman's analysis of variance by ranks**

10.2 Properties

- Additivity - the sum of independent chi-squared variables is also chi-squared distributed.

  Specifically,

  if $\{X_i\}_{i=1}^n$ - are independent chi-squared variables with $\{k_i\}_{i=1}^n$ degrees of freedom, respectively,

  then

  $Y = X_1 + ... + X_n$ - is chi-squared distributed with $k_1 + ... + k_n$ degrees of freedom

- because the chi-squared variable is the sum of squared valued, it can take on only non-negative values and is asymmetric

- as k approaches infinity, the chi-squared distribution converges to normal distribution

10.3 Moments - Summary

$$\mu = k \tag{10.1}$$

$$\sigma^2 = 2k \tag{10.2}$$

$$\gamma_1 = 2\sqrt{\frac{2}{k}} \tag{10.3}$$

$$\gamma_2 = \frac{12}{k} \tag{10.4}$$

Chapter 11

Student t-Distribution

The Student t-distribution with m degrees of freedom is defined to be the distribution of

$$\frac{Z}{\sqrt{\frac{W}{m}}} \tag{11.1}$$

where:

- Z - is a random variable with a standard normal distribution,
- W - is a random variable with a chi-squared distribution with m degrees of freedom
- ans Z and W are independent.

Under the null hypothesis, the t-statistic computed using the homoskedasticity-only standard error could be written in this form.

Chapter 12

F Distribution

If U_1 and U_2 are two independent chi-squared distributions with k_1 and k_2 degrees of freedom, respectively, then X:

$$X = \frac{U_1/k_1}{U_2/k_2} \sim F(k_1, k_2) \tag{12.1}$$

follows an F-distribution with parameters $k1$ and $k2$.

12.1 Properties

- because the chi-squared PDF is zero for negative values, the F-distribution density function is also zero for negative values
- as k_1 and k_2 increase, the mean and mode converge to one
- as k_1 and k_2 approach infinity, the F-distribution converged to a normal distribution
- the square of a variable with a t-distribution has an F-distribution. More

specifically, if X is a random variable with a t-distribution with k degreed of freedom, then X^2 has an F-distribution with 1 and k degrees of freedom:

$$X^2 \sim F(1, k) \tag{12.2}$$

12.2 Construction of F-statistics

12.2.1 Comparison of variances of 2 distributions

Hypothesis:

$$\begin{aligned} H_0 &= \sigma_1^2 = \sigma_2^2 \\ H_1 : &= \sigma_1^2 < \sigma_2^2, \text{ for a lower one-tailed test} \\ H_1 : &= \sigma_1^2 > \sigma_2^2, \text{ for an upper one-tailed test} \\ H_1 : &= \sigma_1^2 \neq \sigma_2^2, \text{ for a two-tailed test} \end{aligned} \tag{12.3}$$

Test:

$$F = \frac{s_1^2}{s_2^2} \tag{12.4}$$

where s_1^2 and s_2^2 are the sample variances.

12.2.2 Joint hypothesis

Chapter 13

Confidence Intervals and Hypothesis Testing

13.1 Sample Mean

The sample mean is a random variable.

If we increase the sample size, our sample mean estimation will be closer to the real mean. The reason is simple: a single outlier will have much less impact.

If our sample size is n and the true variance of our Data Generating Process (DGP) is σ^2, them the variance of the sample mean is:

$$\sigma_\mu^2 = \frac{\sigma^2}{n} \tag{13.1}$$

It follows that the standard deviation of the sample mean decreases with the square root of n.

This is yet another example of the famous square root rule for independent and identically distributed (i.i.d.) variables.

From the central limit theorem, the distribution of the mean converges to the a normal distribution.

13.2 Sample Variance

For a given DGP id we repeatedly calculate the sample variance, the expected value of the sample variance will equal the true variance, and the variance of the sample variance will equal:

$$\mathbb{E}\Big[\big(\hat{\sigma}^2 - \sigma^2\big)^2\Big] \quad = \quad \sigma^4\left(\frac{2}{n-1} + \frac{k_{ex}}{n}\right) \tag{13.2}$$

where:

- n - is the sample size,
- and k_{ex} - is the excess kurtosis

If the GDP has a normal distribution, then we can also say something about the shape of distribution of the sample variance. If we have n sample points and $\hat{\sigma^2}$ is the sample variance, then our estimator will follow a chi-squared distribution with $(n-1)$ degrees of freedom:

$$(n-1)\frac{\hat{\sigma^2}}{\sigma^2} \sim \chi^2_{n-1} \tag{13.3}$$

where σ^2 - is the population variance. Note, that this is true only when the DGP has a normal distribution. Unfortunately, unlike the case of the sample mean we cannot apply the central limit theorem here. Even when the sample size is large, if the underlying distribution is non-normal, the statistic in the equation above can vary significantly from a chi-squared distribution.

13.3 Confidence Intervals

In our discussion of a sample mean, we assumed that the standard deviation of the underlying distribution was known. In practise, the true standard deviation is likely not to be known. At the same time we are measuring the sample mean, we will typically be measuring the sample variance as well.

It turns out, that if we first standardise our estimate of the sample mean using the sample standard deviation, the new random variable follows a Student's t distribution with $(n-1)$ degreed of freedom:

$$t = \frac{\hat{\mu} - \mu}{\hat{\sigma}\sqrt{n}} \tag{13.4}$$

Here the numerator is simply the difference between the sample mean and the population mean, while the denominator is the sample standard deviation divided by the square root of the sample size.

In practise the population mean μ is often unknown.

We often write:

$$P\left[\hat{\mu} - \frac{x_L\hat{\sigma}}{\sqrt{n}} \leq \mu \leq \hat{\mu} + \frac{x_U\hat{\sigma}}{\sqrt{n}}\right] \quad = \quad \gamma \tag{13.5}$$

When it is formulated this way, we call this range the confidence interval for the population mean.

13.4 Hypothesis Testing

In addition to the null hypothesis, we can offer an alternative hypothesis.

In principle, we could test any number of hypothesis. In practice, as long as the

alternative is trivial, we tend to limit ourselves to stating the null hypothesis. It is common to construct null hypothesis so that the desired result is false. We often find that the desired outcome fo a null hypothesis is rejection.

13.4.1 One Tail or Two?

The difference between a one-sided test and a two-side test is that while the alternative hypothesis in former explores the possibility of a change in only one direction (increase or decrease), the latter explores the possibility of a change in either direction.

13.4.2 Type Error

A type 1 error is the incorrect rejection of a true null hypothesis (also known as a "false positive" finding), while a type 2 error is incorrectly retaining a false null hypothesis (also known as a "false negative" finding).
Simpler stated, a type 1 error is to falsely infer the existence fo something that is not there, while a type 2 error is to falsely infer the absence of something that is.

13.5 P-Values

P-value shows the lowest (significance) level at which H_0 can be rejected. The p-value of a result, p, is the probability of obtaining a result at least as extreme, given that the null hypothesis were true.

13.6 Significance Level/The Size of a Test

Significance level (sometimes it is called **"the size of a test"**) α is the probability of the study rejecting the null hypothesis, given that it were true. Decreasing the test significance level, when conducting a hypothesis test, will decrease the likelihood of rejecting the null hypothesis when it is in fact true.

13.7 F-statistic

An F-test is any statistical test in which the test statistic has an F-distribution under the null hypothesis.
Common examples of the use of F-tests include the study of the following cases:

- the hypothesis that the means of a given set of normally distributed populations, all having the same standard deviation, are equal. This is perhaps the best known F-test, and plays an important role in the analysis of variance (ANOVA)
- F-test of the equality of two variances (the F-test is sensitive to normality)
- the hypothesis that a proposed regression model fits the data well
- the hypothesis that a data set in a regression analysis follows the simpler of two proposed linear models that are nested with each other

Most F-tests arise by considering a decomposition of he variability in a collectionof data in terms of sums of squares. The test statistic in an F-test is the ratio of two scaled sums of squares reflecting two sources of variability. These sums of squares are constructed so that the statistic tends to be greater when the null hypothesis is not true. In order for statistics to follow the F-distribution

under the null hypothesis, the sums of squares should be statistically independent, and each should follow a scaled χ^2-distribution. The later condition is guaranteed if the data values are independent and normally distributed with a common variance.

F-statistic are based on the ration of mean squares. The term "mean squares" means an estimate of population variance that accounts for the degrees of freedom used to calculate that estimate.

13.7.1 ANOVA

When comparing the variances of two different populations,we use F-statistic, computed as:

$$F = \frac{S_2^2}{S_2^2} \tag{13.6}$$

where S_1^2 and S_2^2 are the sample variances.

The F-statistic has $(n_1 - 1, n_2 - 1)$ degrees of freedom.

13.8 Important Values to Remember for Normal Distribution

13.8.1 One-Tailed Test

$$
\begin{aligned}
N(-1.282 \text{ or } -1.28) &= 10\% \\
N(-1.645 \text{ or } -1.65) &= 5\% \\
N(-1.96) &= 2.5\% \\
N(-2.326 \text{ or } -2.33) &= 1\% \\
N(-1) &= 15.9\% \text{ because } \pm 1 \text{ standard deviation is } 68\% \text{ area under the curve} \\
N(-2) &= 2.3\% \text{ because } \pm 2 \text{ standard deviations is } 95\% \text{ area under the curve} \\
N(-3) &= 0.3\% \text{ because } \pm 3 \text{ standard deviations is } 99.7\% \text{ area under the curve}
\end{aligned}
$$

13.8.2 Two-Tailed Test

$$
\begin{aligned}
N(-1.645) &= 10\% \\
N(-1.96) &= 5\% \\
N(-2.58) &= 1\%
\end{aligned}
$$

Chapter 14

OLS

14.1 Equation

$$Y_i = \beta_0 + \beta_1 X_i + u_i \tag{14.1}$$

The equation is the **linear regression model with a single regressor**, in which Y is the **dependent variable** and X is the **independent variable** or the **regressor**.

The **intercept** β_0 and the **slope** β_1 are the **coefficients** of the population regression line.

The slope β_1 is the change in Y associated with a unit change in X. The intercept is the value of the population regression line when $X = 0$; it is the point at which the population regression line intercepts the Y axis.

The term u_i is the **error term**. The error term incorporates all of the factors responsible for the difference between the Y_i and the value $\hat{Y}_i$ predicted by the population line. This error term contains all the other factors besides X that

determine the value of the dependent variable Y for a specific observation i.

14.2 OLS Estimator, Predicted Values, and Residuals

The OLS estimators are:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\left(X_i - \bar{X}\right)^2} = \frac{S_{XY}}{S_{XX}} \tag{14.2}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} \tag{14.3}$$

The OLS predicted values $\hat{Y}_i$ and residuals $\hat{u}_i$ are:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_i X_i,\ i = 1, ..., n \tag{14.4}$$

$$\hat{u}_i = Y_i - \hat{Y}_i,\ i = 1, ..., n \tag{14.5}$$

14.3 OLS assumptions

1. The conditional distribution of residuals u_i given regressor X_i has mean of zero
2. The pairs (X_i, Y_i) $i = 1, ..., n$ are independently and identically distributed across observations
3. Large outliers are unlikely

14.4 Properties of OLS estimators

OLS estimator is:

1. unbiased[1]
2. consistent - the estimates will converge upon the true values as the sample size, n, increases
3. has a variance that is inversely proportional to the sample size n
4. has a normal sampling distribution when the sample size is large
5. in addition, under certain conditions the OLS estimator is more efficient than some other candidate estimators. Specifically, if the least squares assumptions hold and if the errors are homoskedastic, then the OLS estimator has the smallest variance of all conditionally unbiased estimators that are linear functions of $Y_1, ..., Y_n$

14.5 Measures of Fit

14.5.1 Explained Sum of Squares (ESS)

ESS is the sum of squared deviations of the predicted values of Y_i, $\hat{Y}_i$ from their average:

$$ESS = \sum_{i=1}^{n} \left(\hat{Y}_i \bar{Y}_i\right)^2 \tag{14.6}$$

[1] Bias of an estimator is the difference between this estimator's expected value and the true values of the parameter being estimated.

14.5.2 Total Sum of Squares (TSS)

TSS is the sum of squared deviations of Y_i from its average:

$$TSS = \sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2 \tag{14.7}$$

14.5.3 Sum of Squares Residuals (SSR)

The sum of squared residuals, or SSR, is the sum of the squared OLS residuals:

$$SSR = \sum_{i=1}^{n} \hat{u}_i^2 \tag{14.8}$$

It can be shown that

$$TSS = ESS + SSR \tag{14.9}$$

14.5.4 The R^2

The R^2 (also called "the coefficient of determination") ranges between 0 and 1 and measures the fraction of the variance of Y_i that is explained by X_i. Write the dependent variable Y_i as the sum of the predicted value $\hat{Y}_i$ plus the residual $\hat{u}_i$:

$$Y_i = \hat{Y}_i + \hat{u}_i \tag{14.10}$$

In this notation, the R^2 is the ratio of the sample variance of $\hat{Y}_i$ to the sample variance of Y_i.

Mathematically, the R^2 can be written as the ratio of the explained sum of squares to the total sum of squares.

The R^2 is the ratio of the explained sum of squares to the total sum of squares:

$$R^2 = \frac{ESS}{TSS} \tag{14.11}$$

Thus the R^2 also can be expressed as 1 minus the ratio of the sum of squared residuals to the total sum of squares:

$$R^2 = 1 - \frac{SSR}{TSS} \tag{14.12}$$

The R^2 of the regression of Y on the single regression X is the square of the correlation coefficient between Y and X.

14.5.5 Standard Error of the Regression (SER)

The standard error of the regression measures how far Y_i typically if from its predicted value.

The standard error of the regression (SER) is an estimation of the standard deviation of the regression error u_i.

The units of u_i and Y_i are the same, so the SER is a measure of the spread of the observations around the regression lines, measured in the units of the dependent variable.

For example, if the units of the dependent variable are dollars, then the SER measures the magnitude of a typical deviation from the regression line - that is, the magnitude of a typical regression error - in dollars.

Because the regression errors $u_1, ..., u_2$ are unobserved, the SER is computed using their sample counterparts, the OLS residuals $\hat{u}_1, ..., \hat{u}_n$. The formula for

the SER is:

$$SER \quad = \quad s_{\hat{u}}, \text{ where } s_{\hat{u}}^2 = \frac{1}{n-2}\sum_{i=1}^{n}\hat{u}_i^2 = \frac{SSR}{n-2} \tag{14.13}$$

where the formula for $s_{\hat{u}}$ uses the fact that the sample average of the OLS residuals is zero.

The formula for the SER is similar to the formula for the sample standard deviation of Y given earlier, except that $Y_i - Y_Y$ is replaced by $\hat{u}_i$, and the divisor is $n-1$, whereas here is is $n-2$. The reason for using the divisor $n-2$ here (instead of n) is the same as the reason for using the divisor $n-1$. It corrects for a slight downward bias introduced because two regression coefficients were estimated.

This is called a **"degrees of freedom correction"**. Because two coefficients were estimated (β_0 and β_1), two "degrees of freedom" of the data were lost, so the divisor in this factor is $n-2$.

14.6 Gauss-Markov theorem

If

- the three least squares assumptions hold
- and if the error is homoskedastic,

then

the OLS estimator has the smallest variance, conditional on $X_1, ..., X_n$ among all estimators in the class of linear conditionally unbiased estimators.

In other words, the OLS estimator is the **B**est **L**inear conditionally **U**nbiased

Estimator - that is, it is **BLUE**.

This result extends to regression the result that the sample average $\bar{Y}$ is the most efficient estimator of the population mean among the class of all estimators that are unbiased and are linear functions (weighed averages) of $Y_1, ..., Y_n$.

14.7 Homoskedastic Normal Regression Assumptions

If the three least squares assumptions hold, and in addition:

- the errors are homoskedastic
- and the errors are normally distributed

These five assumptions are collectively called **homoskedastic normal regression assumptions**

Chapter 15

Regression with a Single Regressor

15.1 Testing Hypothesis About the Population Mean

Recall that the null hypothesis that the mean of Y is a specific value $\mu_{Y,0}$ can be written as $H_0 : E\big(Y\big) = \mu_{Y,0}$ and the two-sided alternative is $H_1 : E\big(Y\big) \neq \mu_{Y,0}$. the test of the null hypothesis H_0 against the two-sided alternative proceeds as:

1. compute the standard error $SE\big(\bar{Y}\big)$ of $\bar{Y}$, which is an estimator of the standard deviation of the sampling distribution of $\bar{Y}$

2. compute the t-statistic, which has the general form

$$t = \frac{\bar{Y} - \mu_{Y,0}}{SE\big(\bar{Y}\big)} \tag{15.1}$$

3. compute p-value, which is the smallest significance level at which the null

hypothesis could be rejected, based on the test statistic actually observed. Because the t-statistic has a standard normal distribution in large samples under the null hypothesis, the p-value for a two-sided hypothesis test is:

$$p = 2\Phi\left(-\left|t^{act}\right|\right) \tag{15.2}$$

where

- t^{act} is the value of the t-statistic actually computed
- and Φ is the cumulative standard normal distribution.

Alternatively, the third step can be replaced by simply comparing the t-statistic to the critical value appropriate for the test with the desired significance level.

15.2 Testing Hypothesis About the Slop β_1

At a theoretical level, the critical feature justifying the foregoing testing procedure for the population mean is that, in large samples, the sampling distribution of $\bar{Y}$ is approximately normal. Because $\hat{\beta}_1$ also has a normal sampling distribution in large samples, hypothesis about the true value of the slope β_1 can be tested using the same general approach.

Under the null hypothesis the true population slope β_1 takes on some specific value $\beta_{1,0}$. Under the two-sided alternative $\beta_1 \neq \beta_{1,0}$. That is the **null hy-**

pothesis and **two-sided alternative hypothesis** are:

$$H_0 : \beta_1 = \beta_{1,0} \text{ vs} \tag{15.3}$$

$$H_1 : \beta_1 \neq \beta_{1,0} \tag{15.4}$$

$$\text{two-sided alternative} \tag{15.5}$$

To test the null hypothesis, we follow the same three steps as for the population mean:

1. compute the **standard error of** $\hat{\beta}_1$
2. compute the ***t*-statistic**:

$$t = \frac{\hat{\beta}_1 - \beta_{1,0}}{SE(\hat{\beta}_{1,0})} \tag{15.6}$$

3. compute the p-value, the probability of observing a value of $\hat{\beta}_1$ at least as different from $\beta_{1,0}$ as the estimate actually computed $\left(\hat{\beta}_1^{act}\right)$, assuming that the null hypothesis is correct.

Because $\hat{\beta}_1$ is approximately normally distributed in large samples, under the null hypothesis the t-statistic is approximately distributed as a standard normal random variable, so in large samples:

$$p = Pr\left(\left|Z\right| > \left|t^{act}\right|\right) = 2\Phi(-\left|t^{act}\right|) \tag{15.7}$$

A small value of the p-value, say less 5%, provides evidence against the null hypothesis in the sense that the chance of obtaining a value of $\hat{\beta}_1$ by pure random variable from one sample to the next is less than 5% if, in fact, the null hypothesis is correct. If so, the null hypothesis is rejected at the 5% significance level.

Alternatively, the hypothesis can be tested at the 5% significance level simply by comparing the value of the t-statistic to the critical value for a two-sided test, and rejecting the null hypothesis at the required level if $\left|t^{act}\right| > \left|t^{critical}\right|$, say $\left|t^{act}\right| > 1.96$

15.2.1 One-Sided Hypotheses Concerning β_1

For a one-sided test, the null hypothesis and the one-sided alternative hypothesis are:

$$H_0 \quad : \quad \beta_1 = \beta_{1,0} \qquad \text{vs} \tag{15.8}$$

$$H_1 \quad : \quad \beta_1 < \beta_{1,0} \qquad \text{one-sided alternative} \tag{15.9}$$

where $\beta_{1,0}$ is the value of β_1 under the null hypothesis and the alternative is that $\beta_1 < \beta_{1,0}$. If the alternative is that β_1 is greater than $\beta_{1,0}$, the inequality in the equation above is reversed.

Because the null hypothesis is the same for a one- and a two-sided hypothesis test, the construction of the t-statistic is the same. The only difference between a one- and two-sided hypothesis test is how you interpret the t-statistic. For the one-sided alternative in the equation above, the null hypothesis is rejected against the one.sided alternative for large negative, but not large positive, values of the t-statistic: Instead of rejecting if $\left|t^{act}\right| > 1.96$, the hypothesis is rejected at the 5% significance level if $t^{act} < -1.645$.

The p-value for a one-sided test is obtained from the cumulative standard normal distribution as:

$$p \quad = \quad Pr\big(Z < t^{act}\big) = \Phi\big(t^{act}\big) \quad p\text{-value, one-sided left-tail test} \tag{15.10}$$

If the alternative hypothesis is that β_1 is greater than $\beta_{1,0}$, the inequalities in the equation above are reversed, so the p-value is right-tail probability, $Pr\big(Z > t^{act}\big)$.

15.2.2 When Should a One-Sided Test Be Used?

In practise, one-sided alternative hypotheses should be used only when there is a clear reason for doing so. This evidence could come from economic theory, prior empirical evidence, or both. However, even if it initially seems that the relevant alternative is one sided, upon reflection this might not necessarily be so.

15.3 Testing Hypothesis About the Intercept β_0

Occasionally, however, the hypothesis concerns the intercept β_0. the general approach to testing the null hypothesis consists of the same steps above, applied to β_0.

15.4 Confidence Interval for a Regression Coefficient

15.4.1 Confidence Interval for β_1

Recall that a 95% **confidence interval for** β_1 has two equivalent definitions:

1. It is the set of values that cannot be rejected using a two-sided hypothesis test with a 5% significance level.

2. It is an interval that has a 95% probability of containing the true value of β_1. That is, in 95% of possible samples that might be drawn, the confidence interval will contain the true value of β_1. Because this interval contains the true value in 95% of all samples, it is said to have a **confidence level** of 95%.

To construct the confidence interval, note that the t-statistic will reject the hypothesized value $\beta_{1,0}$ whenever $\beta_{1,0}$ is outside the range $\hat{\beta}_1 \pm 1.96 \cdot SE(\hat{\beta}_1)$:

$$CI = \hat{\beta}_1 \pm 1.96 \cdot SE(\hat{\beta}_1) \tag{15.11}$$

For small samples the confidence interval is calculated as:

$$CI = \hat{\beta}_i \pm \left(t_{cl,n-2} \cdot SE(\beta_i)\right) \tag{15.12}$$

where $t_{cl,n-2}$ - is the t-distribution for a given confidence level cl for a given number of observations n.

15.5 Heteroskedasticity and Homoskedasticity

15.5.1 Definition

The error term u_i is **homoskedastic** if the variance of the conditional distribution of u_i given X_i is constant for $i = 1, ..., n$ and in particular does not depend on X_i. Otherwise, the error term is **heteroskedastic**.
The definition of homoskedasticity states that the variance of u_i does not depend

on the regressor.

15.5.2 Mathematical Implications of Homoskedasticity

The OLS Estimators Remain Unbiased and Asymptotically Normal

Because the least squared assumptions place no restrictions on the conditional variance, they apply to both the general case of heteroskedasticity and the special case of homoskedasticity.
Therefore,

- the OLS estimators remain unbiased ans consistent even if the error are homoskedastic.
- In addition, the OLS estimators have sampling distributions that are normal in large samples even if the errors are homoskedastic.

Whether the errors are homoskedastic or heteroskedastic, the OLS estimator is

- unbiased,
- consistent,
- and asymptotically normal.

However, the standard errors and statistical inferences must be calculated with heteroskedasticity robust methods.

Practical Implications

The simplest thing is always to use the heteroskedasticity-robust standard errors.
Heteroskedastic data still provides an unbiased estimate, but standard errors - and potentially, statistical inferences - are suspect.

Chapter 16

Multiple Regression

16.1 Omitted Variable Bias

If

- an omitted variable is correlated with an included regressor,
- and the omitted variable is a determinant, in part, of the dependent variable Y,

then the OLS estimator will have **omitted variable bias**.

16.1.1 Omitted Variable Bias and the First Least Squares Assumption

Omitted variable bias means that the first least squared assumption $\mathbf{E}(\mathbf{u_i}|\mathbf{X_i}) = \mathbf{0}$ is incorrect.

To see why, recall that **the error term u_i in the linear regression model**

represents all the factors, other than X_i, that are determinants of Y. If one of those factors is correlated with X_i, this means that the error term (that contains this factor) is correlated with X_i.

In other words, if an omitted variable is a determinant of Y_i, then it is in the error term, and if it is correlated with X_i then the error term is correlated with X_i. **Because u_i and X_i are correlated, the conditional mean of u_i given X_i is non-zero.**

This correlation therefore violates the first least squares assumption and **the OLS is biased.**

This bias does not vanish even in very large samples, and the OLS estimator is inconsistent.

16.1.2 A Formula for Omitted Variable Bias

Let the correlation between X_i and u_i be $corr\,(X_i, u_i) = \rho_{Xu}$.

Suppose that the second and third least squares assumptions hold, but the first does not because ρ_{Xu} is non-zero. Then the OLS estimator has the limit:

$$\hat{\beta}_1 \xrightarrow{\rho} \beta_1 + \rho_{Xu}\frac{\sigma_u}{\sigma_X} \tag{16.1}$$

That is, as the sample size increases, $\hat{\beta}_1$ is close to $\beta_1 + \rho_{Xu}\frac{\sigma_u}{\sigma_X}$ with increasingly high probability.

16.1.3 Summary

1. omitted variable bias is a problem whether the sample size is large or small
2. whether this bias is large or small in practice depends on the correlation

ρ_{Xu} between the regressor and the error term. The larger is $|\rho_{Xu}|$, the larger is the bias.

3. the direction of the bias in $\hat{\beta}_1$ depends on whether X and u are positively or negatively correlated.

16.1.4 Addressing Omitted Variable Bias by Dividing the Data into Groups

One way to address the omitted variable bias is to divide the data into groups.

16.2 Measures of Fit in Multiple Regression

Three commonly used summary statistics in multiple regression are:

1. the standard error of the regression
2. the regression R^2
3. ans the adjusted R^2 (also known as $\bar{R}^2$)

16.2.1 The Standard Error of the Regression (SER)

The SER is a measure of the spread of the distribution of Y around the regression line.

$$SER = s_{\hat{u}} \tag{16.2}$$

where

$$s_{\hat{u}}^2 = \frac{1}{n-k-1}\sum_{i=1}^{n}\hat{u}_i^2 = \frac{SSR}{n-k-1} \tag{16.3}$$

where SSR is the sum of squared residuals,

$$SSR = \sum_{i=1}^{n}\hat{u}_i^2 \tag{16.4}$$

Using $n-k-1$ is called ***degress-of-freedom-adjustment***.

16.2.2 The R^2

The regression R^2 is the fraction of the sample variance of Y_i explained by the regressors. Equivalently, the R^2 is 1 minus the fraction of the variance of Y_i not explained ba the regressors.

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{SSR}{TSS} \tag{16.5}$$

where the explained sum of squares is

$$ESS = \sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}\right)^2 \tag{16.6}$$

and the total sum of squares is

$$TSS = \sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2 \tag{16.7}$$

In multiple regression, the R^2 increases whenever a regressor is added, unless the estimated coefficient on the added regressor is exactly zero.

16.2.3 The "Adjusted R^2 or ($\bar{R}^2$)"

Because the R^2 increases when a new variable is added, an increase in the R^2 does not mean that adding a variable actually improves the fit of the model. The **adjusted** R^2, or $\bar{R}^2$ is a modified version of the R^2 that does not necessarily increase when a new regressor is added. The $\bar{R}^2$ is:

$$\bar{R}^2 = 1 - \frac{n-1}{n-k-1}\frac{SSR}{TSS} = 1 - \frac{s_u^2}{s_Y^2} \tag{16.8}$$

$\bar{R}^2$ properties

1. $$\frac{n-1}{n-k-1} > 1 \Rightarrow \bar{R}^2 < R^2 \tag{16.9}$$

2. adding a regressor has two opposite effects on $\bar{R}^2$. On the one hand, the SSR falls, which increases the $\bar{R}^2$. On the other hand, the factor $(n-1)/(n-k-1)$ increases. Whether the $\bar{R}^2$ increases or decreases depends on which of this two effects is stronger.

3. the $\bar{R}^2$ can be negative.

16.3 The Least Squared Assumptions in Multiple Regression

There are 4 least squares assumptions in the multiple regression model. The first 3 are those for the single regressor model. The fourth assumption is **no perfect multicollinearity**:

1. the conditional distribution of u_i given $X_{1i}, ..., X_{ki}$ has a mean of zero:

$$\mathbb{E}\left(u|X_{1i}, ..., X_{ki}\right) = 0 \tag{16.10}$$

2. $(X_{1i}, ..., X_{ki}, Y_i)$, $i = 1, ..., n$ are $i.i.d$

3. large outliers are unlikely

4. no perfect multicollinearity

The regressors are said to be **perfectly multi-collinear** (or to exhibit **perfect multi-collinearity**) if one of the regressors is a perfect linear combination of other regressors.

The mathematical reason for this failure is that perfect multi-collinearity produces division by zero in the OLS formulas.

At an intuitive level, perfect multi-collinearity is a problem because you are asking the regression to answer an illogical question. In multiple regression, the coefficient on one of the regressors is the effect of a change in that regressor, holding the other regressors constant.

In general, **the solution to perfect multi-collinearity is to modify the regressors to eliminate the problem**.

16.4 Perfect Multi-collinearity

Perfect multi-collinearity typically arises when a mistake has been made in specifying the regression.

16.4.1 Dummy Variable Trap

If

- there are G binary variables,
- if each observation falls into one and only one category,
- if there is an intercept in the regression,
- and if all G binary variables are included as regressors,

then the regression will fail because of perfect mutli-collinearity.
This situation is called the **dummy variable trap**.
The usual way to avoid the dummy variable trap is to exclude one of the binary variables from the multiple regression, so only $G - 1$ of the G binary variables are included as regressors.

16.5 Imperfect Multi-Collinearity

Imperfect multi-collinearity arises when one of the regressors is very highly but not perfectly correlated with the other regressors.
Unlike the perfect multi-collinearity, imperfect multi-collinearity

- does not prevent estimation of the regressions,
- not does it imply a logical problem with the choice of regressors.

However, it does mean that **one or more regression coefficients could be estimated imprecisely**.
The effect of imperfect multi-collinearity on the variance of the OLS estimators

can be seen mathematically by inspecting the variance of $\hat{\beta}_1$ in a multiple regression with two regressors (X_1 and X_2) for the special case of a homoskedastic error:

$$\sigma^2_{\hat{\beta}_1} = \frac{1}{n}\left[\frac{1}{1-\rho^2_{X_1,X_2}}\right]\frac{\sigma^2_u}{\sigma^2_{X_1}} \tag{16.11}$$

In this case, the variance of $\hat{\beta}_1$ is inversely proportional to $1-\rho^2_{X_1,X_2}$, where $\rho^2_{X_1,X_2}$ is the correlation between X_1 and X_2.

The larger is the correlation between the two regressors, the closer is this term to zero and the larger is the variance of $\hat{\beta}_1$.

In contrast to the perfect multi-collinearity, imperfect multi-collinearity is not necessarily an error, but rather just a feature of OLS, your data, and the question you are trying to answer.

16.6 Joint Hypothesis

Joint hypothesis is a hypothesis that imposes two or more restrictions on the regression coefficients.

If **any one or more** of the equalities under the null hypothesis H_0 is false, than the joint hypothesis itself if false. Thus the alternative hypothesis is that **at least one of the equalities in the null hypothesis H_0 does not hold**.

16.6.1 Why Can't Just Test Individual Coefficients One at a Time

This "one at a time" method rejects the null hypothesis too often, because it takes too many chances - if it fails to reject using the first t-statistic, it gets to

try again using the second.

16.6.2 Bonferroni Method

Advantage - it applies very generally.

Disadvantage - it can have low power - it frequently fails to reject the null hypothesis when in fact the alternative hypothesis is true.

16.7 F-statistic

The F-test is only applicable **if and only if there is a linear relationship between two variables**. **It cannot be used in the presence of squares or products between variables**.

For a regression model with k independent variables and n observations, the F-statistic is defined as:

$$F_{k,n-k-1} = \frac{ESS/k}{SSR/\big(n-k-1\big)} \tag{16.12}$$

In joint hypothesis tests, the aim is always to establish the statistical significance of **at least one** of the regression coefficients. **If the computed F-statistic is greater than the 1-tailed F-value at a given significance, at least one of the coefficients is statistically significantly different from zero; otherwise none of the coefficients are statistically significant**.

If F-test on a multiple regression model establishes that a significant amount of variation in the Y variable is explained by the set of X variables, then we should perform t-test on each X variable to establish whether there is any of them, whose effect on Y is not statistically significant.

16.7.1 Q Restrictions

For q restrictions in large samples under the null hypothesis the F-statistics is distributed $F_{q,\infty}$.

16.7.2 The F-Statistics when $q = 1$

When $q = 1$, the F-statistics tests a single restriction. Then the joint null hypothesis reduces to null hypothesis on a single regression coefficient, and the F-statistics is the square of the t-statistic.

16.7.3 The Homoskedasticity-Only F-Statistics

If the error term u_i is homoskedastic, the F-statistic can be written in terms of the improvement in the fit of the regression as measured by the sum of squared residuals or by the regression R^2.

The resulting F-statistics is referred to as the homoskedastic-only F-statistic.

The homoskedasticity-only F-statistic is computed using a simple formula based on the sum of squared residuals from two regressions.

In the first regression, called the **restricted regression**, the null hypothesis is forced to be true. When the null hypothesis is of the type $H_0 : \beta = 0$, where all the hypothesized values are zero, the restricted regression is the regression in which those coefficients are set to zero, that is, the relevant regressors are excluded from the regression.

In the second regression, called the **unrestricted regression**, the alternative hypothesis is allowed to be true.

If the sum of squared residuals is sufficiently smaller in the unrestricted than

the restricted regression, then the test rejects the null hypothesis.

The **homoskedasticity-only F-statistic** is given by the formula:

$$F = \frac{(SSR_{restricted} - SSR_{unrestricted})/q}{SSR_{unrestricted}/(n - k_{unrestricted} - 1)} \tag{16.13}$$

where:

- $SSR_{restricted}$ - is the sum of squared residuals from the restricted regression,
- $SSR_{unrestricted}$ - is the sum of squared residuals from the unrestricted regression,
- q - is the number of restrictions under the null hypothesis,
- and $k_{unrestricted}$ - is the number of regressors in the unrestricted regression.

An alternative equivalent formula for the homoskedasticity-only F-statistic is based on the R^2 of the two regressions:

$$\begin{aligned} F &= \frac{\left(R^2_{unrestricted} - R^2_{restricted}\right)/k}{(1 - R^2_{unrestricted})(n - k_{unrestricted} - 1)} \\ &= \frac{\left(TSS - SSR\right)/k}{SSR/\left(n - k - 1\right)} \\ &= \frac{ESS/k}{SSR/(n - k - 1)} \end{aligned} \tag{16.14}$$

16.8 Pitfalls when using the R^2 or $\bar{R}^2$

1. An increase in the R^2 or $\bar{R}^2$ does not necessarily mean that an added variable is statistically significant. To be able to confidently determine the significance of an added regressor, a t-test must be performed.

2. A high R^2 or $\bar{R}^2$ does not mean that the regressors are a true cause of the dependent variable.

3. A high R^2 or $\bar{R}^2$ does not mean there is no omitted variable bias. Conversely, a low R^2 does not imply that there necessarily is omitted variable bias.

4. A high R^2 or $\bar{R}^2$ does not necessarily mean you have the most appropriate set of regressors, nor does a low R^2 or $\bar{R}^2$ necessarily mean you have an inappropriate set of regressors.

Chapter 17

Information Criteria

Most model selection criteria attempt to find the model with the smallest out-of-sample 1-step-ahead mean squared prediction error (MSE).
The difference among criteria amount to different penalties for the number of degrees of freedom used in estimating the model (that is, the number of parameters estimated). Because all of the criteria are effectively estimates of out-of-sample means square prediction error, they have a negative orientation - the smaller, the better.

17.1 Sum of Squared Residuals (SSR)

SSR is defined as:

$$SSR = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i \right)^2 \tag{17.1}$$

17.2 Total Sum of Squares (TSS)

TSS is defined as:

$$TSS = \sum_{i=1}^{n} \left(Y_i - \bar{Y}_i\right)^2 \tag{17.2}$$

17.3 Explained Sum of Squares (ESS)

ESS is defined as:

$$ESS = \sum_{i=1}^{n} \left(\hat{Y}_i - \bar{Y}\right)^2 \tag{17.3}$$

17.4 R^2

R^2 is defined as:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2}{\sum_{i=1}^{n} \left(Y_i - \bar{Y}_i\right)^2} = 1 - \frac{SSR}{TSS} = \frac{ESS}{TSS} \tag{17.4}$$

17.4.1 R^2 is a Biased Estimator

R^2 is a **biased estimator** based on your sample - it tends to be too high. The magnitude of the bias depends on how many observations are available to fit the model and how many covariates are relative to this sample size. The bias can be particularly large with small sample size and a moderate number of covariates.

17.4.2 Why R^2 Estimator Cannot Be Unbiased

Why is the standard estimate (estimator) of R squared biased? One way of seeing why it can't be unbiased is that by its definition the estimates always lie between 0 and 1. From one perspective this a very appealing property - since the true R squared lies between 0 and 1, having estimates which fall outside this range wouldn't be nice (this can happen for adjusted R squared). However, suppose the true R squared is 0 - i.e. the covariates are completely independent of the outcome Y. In repeated samples, the R squared estimates will be above 0, and their average will therefore be above 0. Since the bias is the difference between the average of the estimates in repeated samples and the true value (0 here), the simple R squared estimator must be positively biased.

Incidentally, lots of estimators we used are not unbiased (more on this in a later post), so lack of unbiasedness doesn't on its own concern is. The problem is that the bias can be large for certain designs where the number of covariates is large relative to the number of observations used to fit the model.

17.5 Mean Squared Error (MSE)

MSE is defined as:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 = \frac{SSR}{n} \tag{17.5}$$

In-sample MSE can't rise when more variable are added to a model, and typically it will fall continuously as more variables are added.

The regression line predicts the average value of y associated with a certain value of x. The MSE (and also RMSE - the root mean squared error) measures

the spread of the y values around that average, hence helps to establish the accuracy of the forecasting model.

MSE is ultimately related to two other diagnostic statistics - the sum of squared residuals (SSR) and R^2.

17.5.1 Data Mining Bias or In-Sample Over-fitting

Analysts avoid the use of in-sample MSE to estimate out-of-sample MSE because in-sample MSE cannot increase even when more variables are incorporated into the forecasting model. Hence, MSE has a downward bias when predicting out-of-sample error variance. This bias is known as **data mining bias** or **in-sample over-fitting**.

17.5.2 Link to SSR

Looking at the MSE formula reveals that the model with the smallest MSE is also the model with smallest sum of squared residuals, because scaling of squared residuals by $1/n$ does not change the ranking. So selecting the model with the smallest MSE is equivalent to selecting the model with the smallest SSR.

17.5.3 Link to R^2

Similarly, recall the formula for R^2:

$$R^2 = 1 - \frac{SSR}{TSS} \tag{17.6}$$

The denominator of the ratio that appears in the formula is just the sum of squared deviations of y from its sample mean (the so called total sum of squared - TSS), which depends only on the data, not on the particular model fit. Thus, selecting the model that minimizes the sum of squared residuals - which is equivalent to selecting the model that minimizes MSE - is also equivalent to selecting the model that maximizes R^2.

The regression model with the highest R^2 is also the one with the lowest MSE.

MSE and R^2 rank regression models the same since neither criterion adjusts for the number of parameters. Both use equal measures of the sum of squared residuals.

17.6 Standard Error of the Regression

To reduce the bias associated with MSE and its relatives, we have to penalize for degrees of freedom used. Thus, let is consider the mean squared error corrected for degrees of freedom:

$$s^2 = \frac{SSR}{n-k-1} \tag{17.7}$$

where:

- k - is the number of degrees of freedom used in model fitting
- and s^2 - is just the usual unbiased estimate of the regression disturbance variance. That is, it is the square of the usual standard error of the regression.

Thus, we define standard error of the regression as:

$$SER = \sqrt{\frac{SSR}{n-k-1}} \tag{17.8}$$

So, selecting the model that minimizes s^2 is also equivalent to selecting the model that minimizes the standard error of the regression.
The unbiased MSE or s^2 should be preferred to R^2 whenever a model is being selected for forecasting purposes.

17.7 R^2 Adjusted for Degrees of Freedom ($\bar{R}^2$)

$\bar{R}^2$ is defined as:

$$\bar{R}^2 = 1 - \frac{n-1}{n-k-1}\frac{SSR}{TSS} = 1 - \frac{s_u^2}{s_Y^2} \tag{17.9}$$

The denominator of the $\bar{R}^2$ expression depends only on the data, not the particular model fit. So the model that minimizes s^2 is also the model that maximizes $\bar{R}^2$.

17.7.1 $\bar{R}^2$ properties

- the adjusted estimator is always less than the standard one
- the larger k (the number of explanatory variables) relative to n (the number of observations), the larger the adjustment
- it is quite possible for $\bar{R}^2$ to be quite negative
- although the $\bar{R}^2$ uses unbiased estimators of the residual variance and the

variance of Y, it is **not unbiased**

17.8 Penalty Factor

To highlight the degree-of-freedom penalty, let's rewrite s^2 as a penalty factor times the MSE:

$$s^2 = \frac{n-1}{n-k-1}\frac{SSR}{n-1} \tag{17.10}$$

As with s^s, many of the most important forecast model selection criteria are of the form "penalty factor times MSE".

17.9 Information Criterion

An information criterion can be defined as a measure of model fit that balances a closer fit to a set of data and an increasing number of parameters.

17.10 Akaike Information Criteria (AIC)

AIC is defined as:

$$AIC = e^{\left(\frac{2k}{n}\right)} \cdot SER^2 \tag{17.11}$$

17.11 Schwarz (or Bayesian) Information Criteria (SIC or BIC)

SIC is defined as:

$$SIC = n^{\left(\frac{k}{n}\right)} \cdot SER^2 \tag{17.12}$$

17.12 Penalty Comparison

How do the penalty factors associated with MSE, s^2, AIC and SIC compare in terms of severity?

All of the penalty factors are functions of k/n, the number of parameters estimated per sample observation. The s^2 penalty is small and rises slowly with k/n. The AIC penalty is a bit larger and still rises slowly with k/n. The SIC penalty is substantially larger and rises at a slightly increasing rate with k/n.

17.13 Model Selection Criteria Consistency

A model selection criteria is consistent, if the following conditions are met

1. when the true model - that is, the **data generating process (DGP)** - is among the model considered, the probability of selecting the true DGP approaches 1 as the sample size gets large

2. when the true model is not among those considered, so that is impossible to select the true DGP, the probability of selecting the best approximation to the true DGP approaches 1 as the sample size gets large.

MSE is inconsistent, because it does not penalize for degrees of freedom, that is why it is unattractive.

s^2 does penalize for degrees of freedom but, as it turns out, not enough to render it a consistent model selection procedure.

The AIC penalizes degrees of freedom more heavily than s^2, but it, too, remains inconsistent; even as the sample size gets large, the AIC selects models that are too large ("overparameterized").

The SIC, which penalizes degrees of freedom most heavily, is consistent.

17.14 Asymptotic Efficiency

Until now we have implicitly assumed, that either the true DGP or the best approximation to the true DGP is in the fixed set of model considered. In that case, SIC is a superior model selection criteria.

However, a potentially more compelling view for forecasters is that both the true DGP and the best approximation to it are much more complicated than any model we fit, in which case we may want to expand the set of models we try as the sample size grows.

An asymptotically efficient model selection criterion chooses a sequence of models, as the sample size get large, whose 1-step-ahead forecast error variances approach the one that would be obtained using the true model with known parameters at a rate at least as fast as that of any other model selection criterion.

The AIC, although inconsistent, is asymptotically efficient, whereas the SIC is not.

17.15 Information Criterion Properties

- AIC always gives model orders that are at least as large as those obtained under the SIC
- if the SSR falls after the addition of an extra term, this does not mean that the value of the information criterion will also fall. If the SSR falls only by a small amount, the information criteria will rise.
- the penalty factors are:

$$n(n-k) \quad \text{for } s^2 \tag{17.13}$$

$$e^{2k/n} \quad \text{for AIC} \tag{17.14}$$

$$n^{k/n} \quad \text{for SIC} \tag{17.15}$$

Chapter 18

Trend

"Trend" refers to a gradual change in the output of a random variable as a result of a tendency of data points of the variable to move in a given direction (positive or negative) over time.

Chapter 19

Seasonality

19.1 The Nature and Source of Seasonality

A seasonal pattern is one that repeats itself every year.

Seasonality arises from links of technologies, preferences, and institutions to the calendar.

19.2 Dealing with Seasonality

One way to deal with seasonality in a series is simply to remove it and then to model and forecast the **seasonally adjusted** or **deseasonalised time series**. This strategy is perhaps appropriate in certain situations, such as when interest centers explicitly on forecasting **non-seasonal fluctuations**.

19.3 Seasonal Model

A key technique for modelling seasonality is **regression on dummy variables**. The pure seasonal model is:

$$y_t = \sum_{i=1}^{s} \gamma_i D_{it} + \epsilon_t \tag{19.1}$$

Effectively, we are just regressing on an intercept, but we allow for a different intercept in each season. Those different intercepts, the γ's are called "seasonal factors"; they summarise the seasonal patterns over the year.

Instead of including a full set of s seasonal dummies, we can include and $s-1$ seasonal dummies and intercept.

In no case, however, should we include s seasonal dummies and an intercept.

19.4 Calendar Effects

- **holiday variation** - refers to the fact that some holiday's dates change over time
- **trading day variation** - refers to the fact that different months contain different numbers of trading or business days.

Chapter 20

Cycles

20.1 (Strictly/Strongly) Stationary Time Series

A time series is stationary ("strictly stationary", "strongly stationary") if its unconditional joint probability does not change when shifted in time. Consequently, parameters such as mean and variance, if they are present, also do not change over time.

20.2 Weak/Covariance Stationary Time Series

The time series is covariance stationary, if

1. its mean is stable over the time:

$$\mathbb{E}\Big(y_t\Big) \quad = \quad \mu \tag{20.1}$$

2. covariance structure be stable over time, that is autocovariance depends

only on the displacement τ and not on time t:

$$\gamma\Big(t,\tau\Big) \quad = \quad \gamma\Big(\tau\Big) \tag{20.2}$$

3. and variance of the series $\Big($or the autocovariance at displacement 0 $\gamma\big(0\big)\Big)$ is finite. It could be shown, that no autocovariance can be larger in absolute value than $\gamma\big(0\big)$, so if $\gamma\big(0\big) < \infty$, then so, too, are all the autocovariances

Stationary autocovariance function is symmetric, that is:

$$\gamma\big(\tau\big) \quad = \quad \gamma\big(-\tau\big) \tag{20.3}$$

20.2.1 Dealing With Covariance-Non-Stationarity

Ways to deal with covariance-non-stationarity:

- use models that give special treatment to non-stationary components such as trend and seasonality, so that the cyclical component that is left over is likely to be covariance stationary
- simple transformations often appear to transform non-stationary series to covariance stationary

20.2.2 Is Covariance Stationarity Too Restrictive?

Although covariance stationarity requires means and covariances to be stable and finite, it places no restrictions on other aspects of the distribution of the series, such as skewness and kurtosis.

In contrast to the unconditional mean and variance, which must be constant by covariance stationarity, the conditional mean and variance need not be constant, and in general we would expect them not to be constant.

20.3 Common Characteristics of Asset Return Distributions

- leptokurtic - more data points are in the tails
- have no trend, either stochastic or deterministic
- lowly correlated

20.4 Autocovariance Function

The autocovariance $\gamma_{t,\tau}$ at displacement/lag τ is just the covariance between y_t and $y_{t-\tau}$:

$$\gamma_{t,\tau} = cov(y_t, y_{t-\tau}) = \mathbb{E}\Big((y_t - \mu)(y_{t-\tau} - \mu)\Big) \tag{20.4}$$

20.5 Autocorrelation Function

The autocorrelation function is obtained by dividing the autocovariance function by the variance.

$$\rho(\tau) = \frac{\gamma(\tau)}{\gamma(0)} \tag{20.5}$$

Note that we always have $\rho_0 = \gamma_0/\gamma_0 = 1$, because any series is perfectly correlated with itself.

20.6 Partial Autocorrelation Function

The partial autocorrelation function $p(\tau)$ is the coefficient of $y_{t-\tau}$ in a population linear regression of y_t on $y_{t-1}, ..., y_{t-\tau}$.

Such regression is called autoregression.

It is easy to see that the autocorrelations and partial autocorrelations, although related, differ in an important way. The autocorrelations are just the "simple" or "regular" correlations between y_t and $y_{t-\tau}$. The partial autocorrelations, on the other hand, measure the association between y_t and $y_{t-\tau}$ after the controlling for the effects of $y_{t-1}, ..., y_{t-\tau+1}$. That is, they measure the partial correlation between y_t and $y_{t-\tau}$.

Also in parallel to the autocorrelation function, the partial autocorrelation at displacement 0 is always 1 and is therefore uninformative and uninteresting.

20.7 White Noise

20.7.1 Definition

Suppose that

$$y_t = \epsilon_t \tag{20.6}$$

$$\epsilon_t \sim \left(0, \sigma^2\right) \tag{20.7}$$

where the "shock" ϵ_t is uncorrelated over time.

We say that ϵ_t and hence y_t is serially uncorrelated.

We assume that $\sigma^2 < \infty$.

Such process, with

- zero mean,
- constant variance,
- and no serial correlation (or autocorrelation is zero except at lag zero, which is actually the variance)

is called **zero-mean white noise**.

If the mean is constant, but not zero, while the other conditions above hold, then such a process is called **white noise**.

By analogy with the white light, which is composed of all colours of the spectrum in equal amounts, we can think of white noise as being composed of a wide variety of cycles of different periodicities in equal amounts.

$$\epsilon_t \sim WN\left(0, \sigma^2\right) \tag{20.8}$$

and hence:

$$y_t \sim WN\left(0, \sigma^2\right) \tag{20.9}$$

20.7.2 Independent White Noise

Note, that although ϵ_t and hence y_t are serially uncorrelated, they are not necessarily serially independent, because they are not necessarily normally distributed.

If in addition to being serially uncorrelated, y **is serially independent**, then we say that y is **independent white noise**. We write:

$$y_t \sim \left(0, \sigma^2\right) \tag{20.10}$$

Another name for independent white noise is **strong white noise**, in contrast to standard serially uncorrelated **weak white noise**.

We say that "y is independently and identically distributed with zero mean and constant variance".

If y is

- serially uncorrelated
- and normally distributed,

then it follows that y is **also independent**, and we say that y is **normal white noise** or **Gaussian white noise**. We write:

$$y_t \sim N\left(0, \sigma^2\right) \tag{20.11}$$

We read "y is independently and identically distributed as normal, with zero mean and constant variance" or simply "y is Gaussian white noise".

20.7.3 White Noise Properties

Recall that the disturbances in a regression model is typically assumed to be white noise of one sort or another.

By construction the unconditional mean of y is:

$$\mathbb{E}\left(y_t\right) = 0 \tag{20.12}$$

and the unconditional variance of y is

$$var\Big(y_t\Big) \quad = \quad \sigma^2 \tag{20.13}$$

Note that unconditional mean and variance are constant. In fact, the unconditional mean and variance must be constant for any covariance stationary process. Because white noise is, by definition, uncorrelated over time, all the autocovariances, and hence all the autocorrelations, are 0 beyond displacement 0.

20.8 Wold's Theorem

Let $\{y_t\}$ be any covariance-stationary process. Then we can write it as:

$$y_t \quad = \quad B\big(L\big)\epsilon_t = \sum_{i=0}^{\infty} b_i \epsilon_{t-i} \tag{20.14}$$

$$\epsilon_t \quad \sim \quad WN\Big(0, \sigma^2\Big) \tag{20.15}$$

where

- $b_0 = 1$
- and $\sum_{i=0}^{\infty} b_i < \infty$

In short, the correct "model" for any covariance stationary series is some infinite distributed lag of white noise, called the **Wold representation**.

The ϵ_t are called **innovations**, because they correspond to the 1-step forecast error that we would make if we were to use a particularly good forecast. That is, the ϵ_t represent that part of the evolution of y that is linearly unpredictable on the basis of the past of y.

Note also that the ϵ_t although uncorrelated are not necessarily independent as

they are not necessarily Gaussian.

20.9 General Linear Process

The process

$$y_t = B(L)\epsilon_t = \sum_{i=0}^{\infty} b_i \epsilon_{t-i} \tag{20.16}$$

$$\epsilon_t \sim WN\left(0, \sigma^2\right) \tag{20.17}$$

where

- $b_0 = 1$
- and $\sum_{i=0}^{\infty} b_i < \infty$

is called the **general linear process**. "General" because any covariance stationary series can be written that way, and "linear" because the Wold representation expresses the series as a linear function of its innovations.

Although Wold's theorem guarantees only serially uncorrelated white noise innovations, we shall sometimes make a stronger assumption of independent white noise innovations to focus the discussion. We do so, for example, in the following characterisation of the conditional moment structure of the general linear process.

Taking means and variances, we obtain the unconditional moments

$$\mathbb{E}(y_t) = \mathbb{E}\left(\sum_{i=0}^{\infty} b_i \epsilon_{t-i}\right) = \sum_{i=0}^{\infty} b_i \mathbb{E}\left(\epsilon_{t-i}\right) = \sum_{i=0}^{\infty} b_i \cdot 0 = 0 \tag{20.18}$$

and

$$\begin{aligned} var\big(y_t\big) &= var\left(\sum_{i=0}^{\infty} b_i \epsilon_{t-i}\right) = \sum_{i=0}^{\infty} b_i^2 var\Big(\epsilon_{t-i}\Big) = \sum_{i=0}^{\infty} b_i \sigma^2 \\ &= \sigma^2 \sum_{i=0}^{\infty} b_i^2 \end{aligned} \tag{20.19}$$

Define the information set $\Omega_{t-1} = \{\epsilon_{t-1}, \epsilon_{t-2}, ...\}$. The conditional mean is:

$$\begin{aligned} \mathbb{E}\Big(y_t|\Omega_{t-1}\Big) &= \mathbb{E}\Big(\epsilon_t|\Omega\Big) + b_1\mathbb{E}\Big(\epsilon_{t-1}|\Omega_{t-1}\Big) + b_2\mathbb{E}\Big(\epsilon_{t-2}|\Omega_{t-1}\Big) + ... \\ &= 0 + b_1\epsilon_{t-1} + b_2\epsilon_{t-2} + ... = \sum_{i=1}^{\infty} b_t \epsilon_{t-i} \end{aligned} \tag{20.20}$$

and the conditional variance is:

$$\begin{aligned} var\Big(y_t|\Omega_{t-1}\Big) &= \mathbb{E}\left(\Big(y_t - \mathbb{E}\Big(y_t|\Omega_{t-1}\Big)\Big)^2|\Omega_{t-1}\right) \\ &= \mathbb{E}\Big(\epsilon_t^2|\Omega_{t-1}\Big) = \mathbb{E}\Big(\epsilon_t^2\Big) = \sigma^2 \end{aligned} \tag{20.21}$$

The key insight is that the conditional mean *moves* over time in response to the evolving information set. The model captures the dynamics of the process, and the evolving conditional mean is one crucial way of summarizing them.

20.10 Analogy Principle

Suppose we know some properties that are satisfied for the "true parameter" in the population. If we can find a parameter value in the sample that causes the sample to mimic the properties of the population, we might use this parameter value to estimate the true parameter.

20.11 Sample mean

$$\bar{y} = \frac{1}{n}\sum_{t=1}^{n} y_t \tag{20.22}$$

20.12 Sample Autocorrelation

The autocorrelation at displacement τ for the covariance stationary series y is:

$$\rho(\tau) = \frac{\mathbb{E}\Big(\big(y_t-\mu\big)\big(y_{t-\tau}-\mu\big)\Big)}{\mathbb{E}\Big(\big(y_t-\mu\big)^2\Big)} \tag{20.23}$$

application of the analogy principle yields a natural estimator:

$$\begin{aligned}\rho(\tau) &= \frac{\frac{1}{n}\sum_{i=\tau+1}^{n}\Big(\big(y_t-\bar{y}\big)\big(y_{t-\tau}-\bar{y}\big)\Big)}{\frac{1}{n}\sum_{i=1}^{n}\Big(y_t-\bar{y}\Big)^2} \\ &= \frac{\sum_{i=\tau+1}^{n}\Big(\big(y_t-\bar{y}\big)\big(y_{t-\tau}-\bar{y}\big)\Big)}{\sum_{i=1}^{n}\Big(y_t-\bar{y}\Big)^2}\end{aligned} \tag{20.24}$$

This estimator, viewed as the function of τ, is called the **sample autocorrelation function** or **correlogram**.

20.12.1 Sample Autocorrelations for White Noise

If a series is white noise, then the distribution of the sample autocorrelations in large sample is:

$$\rho(\tau) \quad \sim \quad N\left(0, \frac{1}{n}\right) \tag{20.25}$$

Their mean is 0, which is to say that the sample autocorrelations are unbiased estimators of the true autocorrelations, which are in fact 0.
The variance of the sample autocorrelations is approximately $\frac{1}{T}$.
Thus, if the series is white noise, then approximately 95% should fall in the interval $\pm 2/\sqrt{n}$

20.13 Box-Pierce Q-statistic

Let

- n - is the total number of observations
- m - is the maximum lag we are considering

We are often interested in whether a series is white noise - that is, whether *all* its autocorrelations are *jointly* 0. Rewrite the expression (20.25):

$$\hat{\rho}(\tau) \quad \sim \quad N\left(0, \frac{1}{n}\right)$$

as

$$\sqrt{n}\hat{\rho}(\tau) \quad \sim \quad N(0,1) \tag{20.26}$$

Squaring both sides yields:

$$n\hat{\rho}(\tau)^2 \quad \sim \quad \chi_1^2 \tag{20.27}$$

It can be shown, that in addition to being approximately normally distributed, the sample autocorrelations at various displacements are approximately independent of one another.

Recalling that **the sum of independent χ^2 variables is also χ^2 with degrees of freedom equal to the sum of the degrees of freedom of the variables summed**, we have shown that the **Box-Pierce Q-statistic**

$$Q_{BP} \quad = \quad n\sum_{\tau=1}^{m}\hat{\rho}(\tau)^2 \tag{20.28}$$

is approximately distributed as a $\chi_m^2(\tau)$ random variable under the null hypothesis that y is white noise.

A slight modification to this, designed to follow more closely the χ^2 distribution in small samples, is:

$$Q_{LB} \quad = \quad n(n+2)\sum_{\tau=1}^{m}\left(\frac{1}{n-\tau}\right)\hat{\rho}^2(\tau) \tag{20.29}$$

under the null hypothesis that y is white noise , Q_{LB} is approximately distributed as a χ_m^2 random variable.

Note that the **Ljung-Box Q-statistic** is the same as the Box-Pierce Q-statistic, except that the sum of squared autocorrelations is replaced by a weighted sum of squared autocorrelations, where the weights are $(n+2)/(n-\tau)$.

For moderate and large n, the weights are approximately 1, so that the Ljung-Box statistic differs little from the Box-Pierce statistic.

Under the H_0 the statistic Q follows a χ_m^2. For significance level α, the critical

region for rejection of the hypothesis of randomness is:

$$Q > \chi^2_{1-\alpha,m} \tag{20.30}$$

where $\chi^2_{1-\alpha,m}$ is the $1-\alpha$-quantile of the chi-squared distribution with m degrees of freedom.

The LB test is commonly used in ARIMA models. Note, that it is applied to the residuals of a fitted ARIMA model, not to the original series, and in such applications the hypothesis actually being tested is that the residuals from the ARIMA model have no autocorrelation. When testing the residuals of an estimated ARIMA model, the degrees of freedom need to be adjusted to reflect the parameter estimation. For example, for an ARIMA(p, 0, q) model, the degrees of freedom should be set to $m - p - q$.

20.13.1 Properties of BP and LB Q-Statistics

- asymptotically (as the sample size increases), the value of the two test statistics will be equal
- BP could be oversized for small samples
- both tests show a tendency to reject the null hypothesis of zero autocorrelations as n tends to infinity
- if the data are white noise, then the BP and LB statistics will both have the same distribution

20.13.2 Selection of m.

Selection of m is done to balance competing criteria. On the one hand, we don't want m too small, because, after all, we are trying to do a joint test on a large part of the autocorrelation function. On the other hand, as m grows relative to n, the quality of the distributional approximation we have invoked deteriorates. In practise, focusing on m in the neighbourhood of $\sqrt{n}$ is often reasonable.

20.14 Sample Partial Autocorrelations

Partial auto-correlations are obtained from linear regression.

If the fitted regression is:

$$\hat{y_t} \quad = \quad \hat{c} + \hat{\beta}_1 y_{t-1} + ... + \hat{\beta}_\tau y_{t-\tau} \tag{20.31}$$

then the **sample partial autocorrelation** at displacement τ is:

$$\hat{\rho}(\tau) \quad = \quad \hat{\beta}_\tau \tag{20.32}$$

Distribution results identical to those we discussed for the sample autocorrelations hold as well for the sample *partial* autocorrelations. That is, if the series is white noise, approximately 95% of the sample partial autocorrelations should fall in the interval $\pm 2/\sqrt{n}$

Chapter 21

MA Models

The moving average model specifies that the output variable depends linearly on the current and lagged unobservable shocks.

The first-order moving average process, or MA(1) process, is:

$$y_t = \epsilon_t + \theta\epsilon_{t-1} = \big(1 + \theta L\big)\epsilon_t \tag{21.1}$$

$$\epsilon_t \sim WN\Big(0, \sigma^2\Big) \tag{21.2}$$

The structure of the MA(1) process, in which only the first lag of the shock appears on the right, forces it to have a very short memory and hence weak dynamics, regardless of the parameter value.

21.1 Unconditional Mean And Variance

The unconditional mean and variance are:

$$\mathbb{E}\big(y_t\big) = \mathbb{E}\big(\epsilon\big) + \theta\mathbb{E}\big(\epsilon_{t-1}\big) = 0 \tag{21.3}$$

and

$$\begin{aligned} var\big(y_t\big) &= var\big(\epsilon_t\big) + \theta^2 var\big(\epsilon_{t-1}\big) && (21.4)\\ &= \sigma^2 + \theta^2\sigma^2 = \sigma^2\big(1+\theta^2\big) && (21.5)\end{aligned}$$

Note, that for a fixed value of σ, as θ increases in absolute values, so, too, does the unconditional variance.

21.2 Conditional Mean and Variance

The conditional mean and variance of an MA(1), where the conditioning information set is

$$\Omega_{t-1} = \{\epsilon_{t-1}, \epsilon_{t-2}, ...\} \tag{21.6}$$

are

$$\begin{aligned} \mathbb{E}\Big(y|\Omega_{t-1}\Big) &= \mathbb{E}\Big(\epsilon_t + \theta\epsilon_{t-1}|\Omega_{t-1}\Big) \\ &= \mathbb{E}\Big(\epsilon_t|\Omega_{t-1}\Big) + \theta\mathbb{E}\Big(\epsilon_{t-1}|\Omega_{t-1}\Big) = \theta\epsilon_{t-1} && (21.7)\end{aligned}$$

and

$$\begin{aligned} var\Big(y_t|\Omega_{t-1}\Big) &= \mathbb{E}\Bigg(\Big(y_t - \mathbb{E}\Big(y_t|\Omega_{t-1}\Big)\Big)^2|\Omega_{t-1}\Bigg) \\ &= \mathbb{E}\Big(y_t^2|\Omega_{t-1}\Big) - \mathbb{E}\Bigg(\mathbb{E}\Big(y_t|\Omega_{t-1}\Big)\Bigg)^2 \\ &= \mathbb{E}\Big(y_t^2|\Omega_{t-1}\Big) = \mathbb{E}\Big(\epsilon_t^2|\Omega_{t-1}\Big) = \mathbb{E}\Big(\epsilon_t^2\Big) = \sigma^2 \end{aligned} \tag{21.8}$$

The conditional mean explicitly adapts to the information set, in contrast to the unconditional mean, which is constant.

Note, however, that only the first lag of the shock enters the conditional mean - more distant shocks have no effect on the current conditional expectation. This is indicative of the one-period memory of MA(1) processes.

21.3 Autocorrelation Function

To compute the autocorrelation function for the MA(1) process, we must first compute the autocovariance function. We have:

$$\begin{aligned} \gamma(\tau) &= \mathbb{E}(y_t, y_{t-1}) = \mathbb{E}\Bigg(\Big(\epsilon_t + \theta\epsilon_{t-1}\Big)\Big(\epsilon_{t-\tau} + \theta\epsilon_{t-\tau-1}\Big)\Bigg) \\ &= \begin{cases} \theta\sigma^2, & \text{if } \tau = 1 \\ 0, & otherwise \end{cases} \end{aligned} \tag{21.9}$$

The autocorrelation function is just the autocovariance function scaled by the variance

$$\rho(\tau) = \frac{\gamma(\tau)}{\gamma(0)} = \begin{cases} \frac{\theta}{1+\theta^2}, & \text{if } \tau = 1 \\ 0, & \text{otherwise} \end{cases} \tag{21.10}$$

The key feature here is the sharp *cutoff in the autocorrelation function.* All autocorrelations are 0 beyond displacement 1, the order of the MA process.
Note that the requirements of covariance stationarity (constant unconditional mean, constant and finite unconditional variance, autocorrelation dependent only on displacement) are met for any MA(1) process, regardless of the values of its parameters.

21.4 Autoregressive Representation

If $|\theta| < 1$, then we say that the MA(1) process is **invertible**.
In that case, we can "invert" the MA(1) process and express the current value of the series not in terms of a current shock and a lagged shock, but rather in terms of a current shock and lagged values of the series.
That is called an **autoregressive representation**.
An autoregressive representation has a current shock and lagged observable values of the series on the right, whereas a moving average representation has a current shock and lagged unobservable shocks on the right.
Let us compute the autoregressive representation. The process is:

$$y_t = \epsilon_t + \theta\epsilon_{t-1} \tag{21.11}$$

$$\epsilon_t \sim WN\left(0, \sigma^2\right) \tag{21.12}$$

Thus we can solve for the innovation as

$$\epsilon_t = y_t - \theta\epsilon_{t-1} \tag{21.13}$$

Lagging by successfully more periods gives expressions for the innovations at various dates:

$$\epsilon_{t-1} = y_{t-1} - \theta\epsilon_{t-2} \tag{21.14}$$

$$\epsilon_{t-2} = y_{t-2} - \theta\epsilon_{t-3} \tag{21.15}$$

$$... \tag{21.16}$$

Making use of these expressions for lagged innovations, we can substitute backward in the MA(1) process, yielding:

$$y_t = \epsilon_t + \theta y_{t-1} - \theta^2 y_{t-2} + \theta^3 y_{t-3} - ... \tag{21.17}$$

In lag operator notation, we write the infinite autoregressive representation as

$$\frac{1}{1+\theta L} y_t = \epsilon_t \tag{21.18}$$

Note that the back substitution used to obtain the autoregressive representation only makes sense, and in fact a convergent autoregressive representation only exists, if $|\theta| < 1$ because in the back substitution we raise θ to progressively higher powers.

We can restate the invertibility condition in another way:

The inverse of the root of the moving average lag operator polynomial $1 + \theta L$ *must be less that 1 in absolute value.*

Recall that a polynomial of degree m has m roots. Thus, the MA(1) lag operator polynomial has one root, which is the solution to

$$1 + \theta L = 0 \tag{21.19}$$

The root is $L = -1/\theta$, so its inverse will be less than 1 in absolute value if $|\theta| < 1$, and the two invertibility conditions are equivalent.

Autoregressive representations are appealing to forecasters, because one way or another, if a model is to be used for real-world forecasting, it must link the present observables to the past history of observables, so that we can extrapolate to form a forecast of future observables based on present and past observables. Superficially, moving average models don't seem to meet that requirement, because the current value of series is expressed in terms of current and lagged unobservable shocks, not observable variables.

Finally let us consider the partial autocorrelation function for the MA(1) process. From the infinite autoregressive representation of the MA(1) process, we that the partial autocorrelation function will decay gradually to 0. As we discussed earlier, the partial autocorrelations are just the coefficients on the last included lad in a sequence of progressively higher-order autoregressive approximations. If $\theta > 0$, then the pattern of decay will be one of damped oscillation; otherwise, the decay will be one sided.

Note, however, that the partial autocorrelations are not the successive coefficients in the infinite autoregressive representation. Rather, they are the coefficients of the last included lag in sequence of progressively longer autoregressions. The two are related but distinct.

21.5 The MA(q) Process

Let us consider the general MA(q) process:

$$y_t = \epsilon_t + \theta_1\epsilon_{t-1} + \theta_2\epsilon_{t-2} + ... + \theta_q\epsilon_{t-q} = \Theta\big(L\big)\epsilon_t \tag{21.20}$$

$$\epsilon_t \sim WN\Big(0, \sigma^2\Big) \tag{21.21}$$

where

$$\Theta\big(L\big) = 1 + \theta_1 L + ...\theta_q L^q \tag{21.22}$$

is a q-th order lag operator polynomial.

21.5.1 Properties of MA(q) Process

The properties of the MA(q) processes parallel to those of the MA(1) process in all respects.

1. covariance stationarity
2. it is invertible only if the root condition is satisfied
3. the conditional mean of the MA(q) process evolves with the information set, in contrast to the unconditional moments, that are fixed

Just as the MA(1) process is covariance stationary for any value of its parameters, so, too, is the finite-order MA(q) process.
As with the MA(1) process, the MA(q) process is invertible only if a root condition is satisfied. The MA(q) lag operator polynomial has "q" roots. When $q > 1$, the possibility of complex roots arises. The **condition for invert-**

ibility for invertibility of the MA(q) process is that the inverse of all of the roots must be inside the unit circle, in which case we have the convergent autoregressive representation:

$$\frac{1}{\Theta(L)} y_t = \epsilon_t \tag{21.23}$$

The conditional mean of the MA(q) process evolves with the information set, in contrast to the unconditional moments, that are fixed. In contrast to the MA(1) process, in which the conditional mean depends on only the first lag of the innovation, in the MA(q) case the conditional mean depends on q lags of the innovation. Thus, the MA(q) process has longer memory.

The potentially longer memory of the MA(q) process emerges clearly in its autocorrelation function. In the MA(1) case, all autocorrelations beyond displacement 1 are 0; in the MA(q) case, all autocorrelations beyond displacement q are 0. ***This autocorrelation cutoff is a distinctive property of MA processes.***

The partial autocorrelation function of the MA(q) process, in contrast, decays gradually, in accord with the infinite autoregressive representation, in either oscillating or a one-sided fashion, depending on the parameters of the process.

Chapter 22

AR Models

The autoregressive process is also a natural approximation to the Wold representation. We have seen that under certain conditions a MA process has an autoregressive representation, so an autoregressive process is in a sense the same as a moving average process.

Like the MA process, the AR process has direct motivation; it is simply a stochastic difference equation, a simple mathematical model in which the current value of a series is linearly related to its past values, plus an additive stochastic shock.

22.1 The AR(1) Process

The AR(1) process is:

$$y_t = \phi y_{t-1} + \epsilon_t \tag{22.1}$$

$$\epsilon_t = WN\left(0, \sigma^2\right) \tag{22.2}$$

In lag operator form we write:

$$\big(1-\phi L\big)y_t \quad = \quad \epsilon_t \tag{22.3}$$

The AR(1) process shows more sensitivity to the autoregression coefficient than MA(1), which has a very short memory regardless of parameter value. Thus, the AR(1) model is capable of capturing much more persistent value thank is the MA(1).

Recall that a finite-order MA process is always covariance-stationary but that certain conditions must be satisfied for invertibility, in which case an autoregressive representation exists.

For AR processes the situation is precisely reverse. Autoregressive processes are always invertible - in fact, invertibility is not even an issue, as finite-order autoregressive processes are already are in autoregressive form - but certain conditions must be satisfied for an autoregressive process to be a covariance stationary.

If we begin with the AR(1) process

$$y_t \quad = \quad \phi y_{t-1} + \epsilon_t \tag{22.4}$$

and substitute backward for lagged y-s on the right side, we obtain:

$$y_t \quad = \quad \epsilon_t + \phi\epsilon_{t-1} + \phi^2\epsilon_{t-2} + ... \tag{22.5}$$

In lag operator form we write

$$y_t \quad = \quad \frac{1}{1-\phi L}\epsilon_t \tag{22.6}$$

The MA representation for y-s convergent if and only if $|\phi| < 1$. Thus, $|\phi| < 1$ is the condition for covariance stationarity in the AR(1) case. Equivalently, the condition for covariance stationarity is that the inverse of the root of the autoregressive lag operator polynomial be less than 1 in absolute value.

22.1.1 Unconditional Mean and Variance

From the MA representation of the covariance stationary AR(1) process, we can compute the unconditional mean and variance:

$$\begin{aligned} \mathbb{E}\big(y_t\big) &= \mathbb{E}\big(\epsilon_t + \phi\epsilon_{t-1} + \phi^2\epsilon_{t-2} + ...\big) \\ &= \mathbb{E}\big(\epsilon_t\big) + \phi\mathbb{E}\big(\epsilon_{t-1}\big) + \phi^2\mathbb{E}\big(\epsilon_{t-2}\big) + ... = 0 \end{aligned} \tag{22.7}$$

and

$$\begin{aligned} var\big(y_t\big) &= var\big(\epsilon + \phi\epsilon_{t-1} + \phi^2\epsilon_{t-2} + ...\big) \\ &= \sigma^2 + \phi^2\sigma^2 + \phi^4\sigma^2 + ... = \sigma^2\sum_{i=0}^{\infty}\phi^{2i} \\ &= \frac{\sigma^2}{1-\phi^2} \end{aligned} \tag{22.8}$$

22.1.2 Conditional Moments

The conditional moments, in contrast, are:

$$\begin{aligned} \mathbb{E}\Big(y_t|y_{t-1}\Big) &= \mathbb{E}\Big(\phi y_{t-1} + \epsilon_t|y_{t-1}\Big) \\ &= \phi\mathbb{E}\Big(y_{t-1}|y_{t-1}\Big) + \mathbb{E}\Big(\epsilon_t|y_{t-1}\Big) \\ &= \phi y_{t-1} + 0 = \phi y_{t-1} \end{aligned} \tag{22.9}$$

and

$$\begin{aligned} var\Big(y_t|y_{t-1}\Big) &= var\Big(\phi y_{t-1}+\epsilon_t|y_{t-1}\Big) \\ &= \phi^2 var\Big(y_{t-1}|y_{t-1}\Big)+var\Big(\epsilon_t|y_{t-1}\Big) \\ &= 0+\sigma^2=\sigma^2 \end{aligned} \tag{22.10}$$

Note in particular the simple way in which the conditional mean adapts to the changing information set as the process evolves.

22.1.3 Autocorrelation Function

The process is

$$y_t = \phi y_{t-1}+\epsilon_t \tag{22.11}$$

so that, multiplying both sides of the equation by $y_{t-1\tau}$ we obtain

$$y_t y_{t-\tau} = \phi y_{t-1}y_{t-\tau}+\epsilon_t y_{t-\tau} \tag{22.12}$$

For $\tau \geq 1$, taking expectations of both sides gives

$$\gamma(\tau) = \phi\gamma(\tau-1) \tag{22.13}$$

This is called **Yule-Walker equation**.

It is a recursive equation; that is, given $\gamma(\tau)$ for any τ, the Yule-Walker equation immediately tells us how to get $\gamma(\tau+1)$. Thus, if we know $\gamma(0)$, we could use the Yule-Walker equation to determine the entire autocovariance sequence. We do know $\gamma(0)$ - it is just the variance of the process, which we already showed

to be $\gamma(0) = \frac{\sigma^2}{1-\phi^2}$. Thus, we have:

$$\begin{aligned}
\gamma(0) &= \frac{\sigma^2}{1-\phi^2} \\
\gamma(1) &= \phi\frac{\sigma^2}{1-\phi^2} \\
\gamma(2) &= \phi^2\frac{\sigma^2}{1-\phi^2} \\
&\ldots
\end{aligned}$$

and so on.

In general, then

$$\gamma(\tau) = \phi^\tau \frac{\sigma^2}{1-\phi^2}, \tau = 0, 1, 2, ... \tag{22.14}$$

Dividing through by $\gamma(0)$ gives us the autocorrelations:

$$\rho(\tau) = \phi^\tau, \tau = 0, 1, 2, ... \tag{22.15}$$

Note the gradual autocorrelation decay, which is typical of AR processes. The autocorrelations approach 0, but only in the limit as the displacement approaches infinity. In particular, they don't cut off to 0, as is the case for MA processes.

If ϕ is positive, the autocorrelation decay is one-sided. If ϕ is negative, the decay involves back-and-forth oscillations.

22.1.4 Partial autocorrelation function

Finally, the partial autocorrelation function for the AR(1) process cuts off abruptly, specifically:

$$\rho(\tau) = \begin{cases} \phi, & \tau = 1 \\ 0, & \tau > 1 \end{cases} \tag{22.16}$$

It is easy to see why. The partial autocorrelations are just the last coefficients in a sequence of successively longer population autoregressions.

22.2 AR(p) Process

AR(p) is:

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + ... + \phi_q y_{t-p} + \epsilon_t \tag{22.17}$$

In lag operator form, we write:

$$\Phi(L) = \Big(1 - \phi_1 L - \phi_2 L^2 - ... - \phi_p L^q\Big) y_t = \epsilon_t \tag{22.18}$$

22.2.1 Properties

- an AR(p) process is covariance stationary, if and only if the inverses of all roots of the autoregressive lag operator polynomial $\Phi(L)$ are inside the unit circle.
- **necessary condition for covariance stationarity**, which is often useful

as a quick check, is:

$$\sum_{i=1}^{p} \phi_i < 1 \tag{22.19}$$

If the condition is satisfied, the process may or may not be stationary; but if the condition is violated, the process can't be stationary.

- in the covariance stationary case we can write the process in the convergent infinite MA form:

$$y_t = \frac{1}{\Phi(L)}\epsilon_t \tag{22.20}$$

- the autocorrelation function for the general AR(p) process, as with that of AR(1) process, decays gradually with displacement.

- the AR(p) partial autocorrelation function has a sharp cutoff at displacement p, for the same reason that the AR(1) partial autocorrelation function has a sharp cutoff at the displacement 1.

- in spite of the fact, that the qualitative behaviour (gradual damping) of the autocorrelation function matches that of the AR(1), it can nevertheless display a richer variety of patterns, depending on the order and parameters of the process. It can, for example, have damped monotonic decay, as in the AR(1) case with a positive coefficient, but it can also have damped oscillation in ways that AR(1) can't have. In the AR(1) case, the only possible oscillation occurs when the coefficient is negative, in which case the autocorrelations switch signs at each successively longer displacement. In higher-order autoregressive models, however, the autocorrelations can oscillate with much richer patterns reminiscent of cycles in the more traditional sense. This occurs when some roots of the autoregressive lag

operator polynomial are complex.

Chapter 23

ARMA Models

The **AR** part of ARMA indicates that **the evolving variable of interest is regressed on its own lagged values**.

The **MA** part indicates that **the regression error is actually a linear combination of error terms** whose values occurred contemporaneously and at various times in the past. ARMA(p, q) process also have direct motivation.

First, if the random shock that drives an autoregressive process is itself a MA process, then it can be shown that we obtain an ARMA process.

Second, ARMA processes can arise from aggregation. For example, sums of AR processes, or sums or AR and MA processes, can be shown to be ARMA processes.

Finally, AR processes observed subject to measurement error also turn out to be ARMA processes.

Both stationarity and invertibility need to be checked in the ARMA case, because both AR and MA components are present.

23.1 ARMA(1, 1) Process

The simplest ARMA process that is not a pure AR or MA process is the ARMA(1,1) given by:

$$y_t = \phi y_{t-1} + \epsilon_t + \theta \epsilon_{t-1} \tag{23.1}$$

$$\epsilon_t = WN\Big(0, \sigma^2\Big) \tag{23.2}$$

or in lag operator form:

$$(1 - \phi L) y_t = (1 + \theta L) \epsilon_t \tag{23.3}$$

where:

- for stationarity

$$|\phi| < 1 \tag{23.4}$$

  is required

- for invertibility

$$|\theta| < 1 \tag{23.5}$$

  is required

If the covariance stationarity condition is satisfied, then we have the MA representation:

$$y_t = \frac{1 + \theta L}{1 - \phi L} \epsilon_t \tag{23.6}$$

which is an infinite distributed lag of current and past innovations.

Similarly, if the invertibility condition is satisfied, we have the infinite AR representation:

$$\frac{1-\phi L}{1+\theta L} y_t = \epsilon_t \tag{23.7}$$

23.2 ARMA(p, q) Process

The ARMA(p, q) process is a natural generalisation of the ARMA(1, 1) that allows for multiple MA and AR lags. We write:

$$y_t = \phi_1 y_{t-1} + ... + \phi_p y_{t-p} + \epsilon_t + \theta_1 \epsilon_{t-1} + ... + \theta_q \epsilon_{t-q} \tag{23.8}$$

$$\epsilon_t \sim WN\left(0, \sigma^2\right) \tag{23.9}$$

or

$$\Phi\big(L\big) = \Theta\big(L\big)\epsilon_t \tag{23.10}$$

where

$$\Phi\big(L\big) = 1 - \phi_1 L - \phi_2 L^2 - ... - \phi_p L^p \tag{23.11}$$

and

$$\Theta\big(L\big) = 1 + \theta_1 L + \theta_2 L^2 + ... + \Theta_q L^q \tag{23.12}$$

If the inverses of all roots of $\Phi(L)$ are inside the unit circle, then the process is covariance stationary and the convergent infinite MA representation:

$$y_t = \frac{\Theta(L)}{\Phi(L)}\epsilon_t \tag{23.13}$$

If the inverses of all roots of $\Theta(L)$ are inside of the unit circle, then the process is invertible and has convergent infinite AR representation:

$$\frac{\Phi(L)}{\Theta(L)}y_t = \epsilon_t \tag{23.14}$$

23.2.1 Properties

- As with AR and MA, ARMA processes have a fixed unconditional mean but a time-varying conditional mean.
- In contrast to pure MA or AR processes, however, neither the autocorrelation nor partial autocorrelation functions of ARMA processes cut off at any particular displacement. Instead, each damps gradually, with the precise pattern depending on the process.
- ARMA models approximate the Wold representation by a ratio of two finite-order lag operator polynomials, neither of which is degenerate. Thus, ARMA models use ratio of full-fledged polynomials in the lag operator to approximate the Wold representation

$$y_t = \frac{\Theta(L)}{\Phi(L)}\epsilon_t \tag{23.15}$$

23.3 Partial Autocorrelation Function (PACF)

The most important use of the autocorrelation function is in differentiating between AR and ARMA processes. For AR process, the pacf would be zero after p lags. For the ARMA process, the decline in pacf would assume a geometric form.

23.4 Unit Root Test

If one of the time series hat unit root, the linear regression should not be used as the error term of the regression would NOT be covariance stationary, leading to erroneous results.

Chapter 24

ARIMA Models

ARIMA models are applied in some cases where data show evidence of non-stationarity, where an initial differencing step (corresponding to the "integrated" part of the model) can be applied one or more times to eliminate the non-stationarity.

Chapter 25

Volatility

A variable's **volatility** σ is **the standard deviation of the return** provided by the variable per unit of time when the return is expressed using continuous compounding.

If we assume that the returns are independent with the same variance, the variance of the return over T days is T times the variance of the return over one day.

This means that the standard deviation of the return over T days is $\sqrt{T}$ times the standard deviation of the return over one day. This is consistent with the adage "uncertainty increases with square root of time".

25.1 Variance Rate

The variance rate is defined as the square of the volatility.

25.2 Returns

When returns are continuously compounded, the return over many days is the sum of the returns over each of the days:

$$r_{1,\ldots,n} = \sum_{i=1}^{n} r_i \tag{25.1}$$

In calculating volatility the following simplifications are often made:

- **continuous return is replaced with simple return**:

$$u_i = \frac{S_i - S_{i-1}}{S_{i-1}} \tag{25.2}$$

- $\bar{u}$ **is assumed to be zero**.

  The justification for this is that the expected change in a variable in one day is very small when compared with the standard deviation of changes. This is likely to be the case even if he variable happened to increase or decrease quite fast during the m days of our data.

- $m-1$ **in the denominator is replaced with** m.

  This moves us form an unbiased estimate of the volatility to a maximum likelihood estimate.

Under those changes the variance rate is calculated as:

$$\sigma_n^2 = \frac{1}{m}\sum_{i=1}^{m} u_{n-1}^2 \tag{25.3}$$

25.3 The Power Law

The power law provides an alternative to assuming normal distributions. The law asserts that for a variable v

$$Prob(v > x) \quad = \quad Kx^{-\alpha} \tag{25.4}$$

where:

- K and α are constants
- x is large

25.4 Weighting Schemes

Please, note - all the volatility models are applied to variances.
This allows us to apply them later without changes to covariances.
Our objective is to estimate σ_n, the volatility on day n. It therefore makes sense to give more weight to recent data. A model that does this is:

$$\sigma_n^2 \quad = \quad \sum_{i=1}^{m} \alpha_i u_{n-i}^2 \tag{25.5}$$

The weights must sum to unity, so that:

$$\sum_{i}^{m} \alpha_i \quad = \quad 1 \tag{25.6}$$

An extension of the idea above is to assume that there is a long-run average variance rate and that this should be given some weight. This leads to the

model that takes the form:

$$\sigma_n^2 = \gamma V_L + \sum_{i=1}^{m} \alpha_i u_{n-i}^2 \tag{25.7}$$

where V_L is the long run variance rate and γ is the weight assigned to it. Because the weights must sum to unity, we have:

$$\gamma + \sum_{i=1}^{m} \alpha_i = 1 \tag{25.8}$$

This is known as an ARCH(m) model.

Defining

$$\omega = \gamma V_L \tag{25.9}$$

the variance rate can be written

$$\sigma_n^2 = \omega + \sum_{i=1}^{m} \alpha_i u_{n-i}^2 \tag{25.10}$$

where m is the maximum lag.

25.5 Exponentially Weighted Moving Average Model

Let us assume:

$$\alpha_{i+1} = \lambda \alpha_i \tag{25.11}$$

where:

$$0 < \lambda < 1 \tag{25.12}$$

This weighting scheme leads to a particularly simple formula for updating **variance** estimates. The formula is:

$$\sigma_n^2 = \lambda\sigma_{n-1}^2 + (1-\lambda)u_{n-1}^2 \tag{25.13}$$

The lower the value of λ,

- the faster the rate at which old values are 'forgotten'
- and the more weight is given to the most recent observations.

25.6 The GARCH(1, 1) Model

The **variance** is evaluated in the form:

$$\sigma_n^2 = \gamma V_L + \alpha u_{n-1}^2 + \beta\sigma_{n-1}^2 \tag{25.14}$$

where

- γ - is the V_L weight
- α - is the u_{n-1}^2 weight
- β - is the σ_{n-1}^2 weight

Because the weights must sum to one:

$$\gamma + \alpha + \beta = 1 \tag{25.15}$$

The EWMA model is a particular case of GARCH(1, 1) where $\gamma = 0$, $\alpha = 1 - \lambda$ and $\beta = \lambda$.

The "(1, 1)" in GARCH(1, 1) indicates that σ_n^2 is based on the most recent observation of u^2 and the most recent estimate of the variance rate. The more general GARCH(p, q) model calculates σ_n^2 from the most recent p observations on u^2 and the most recent q estimates of the variance rate.

Setting

$$\omega = \gamma V_L \tag{25.16}$$

the GARCH(1, 1) model can also be written

$$\sigma_n^2 = \omega + \alpha u_{n-1}^2 + \beta \sigma_{n-1}^2 \tag{25.17}$$

We also require

$$\alpha + \beta < 1 \tag{25.18}$$

The GARCH(1, 1) model is the same as the EWMA model except that, in addition to assigning weights that decline exponentially to past u_i^2, it also assigns some weight to the long-run average variance rate.

The GARCH(1, 1) model incorporates mean- reversion, whereas EWMA does not.

25.6.1 Persistence

Persistence is defined as:

$$P = \alpha + \beta \tag{25.19}$$

The model with the highest persistence takes the longest to revert to its mean.

25.6.2 Long-run mean variance

If we are given an equation:

$$\sigma_n^2 = \omega + \sum_{i=1}^{p} \alpha_i u_{n-i}^2 + \sum_{j=1}^{q} \beta_{n-j} \sigma_{n-j}^2 \tag{25.20}$$

As

$$\gamma = 1 - \sum_{i=1}^{\max(p,q)} (\alpha_i + \beta_i) \tag{25.21}$$

and

$$\omega = \gamma V_L \tag{25.22}$$

then the GARCH **long-run average variance** is:

$$V_L = \frac{\omega}{1 - \sum_{i=1}^{max(p,q)} (\alpha_i + \beta_i)} \tag{25.23}$$

Chapter 26

Correlation

26.1 Covariance and Correlation

The coefficient of correlation ρ, between two variables V_1 and V_2 is defined as

$$\rho = \frac{\mathbb{E}(V_1V_2) - \mathbb{E}(V_1)\mathbb{E}(V_2)}{SD(V_1)SD(V_2)} \tag{26.1}$$

26.2 Correlation vs Dependence

Two variables are defined as statistically independent if knowledge about one of them does not affect the probability distribution for the other.

Formally, V_1 and V_2 are independent if

$$f(V_2|V_1 = x) = f(V_2) \tag{26.2}$$

where $f(.)$ denotes the probability density function.

The correlation coefficient **measures only linear dependence**.

Another aspect of the way in which V_2 depends on V_1 is found by examining the standard deviation of V_2 conditional on V_1. As we will see later, this is constant when V_1 and V_2 have a bivariate normal distribution. But, in other situations, the standard deviation of V_2 is expected to depend on the value of V_1.

26.3 Monitoring Correlation

It is common to assume, that the expected daily returns are zero when the variance/covariances rates per day are calculated.
This means that the covariance rate per day between X and Y on day n is assumed to be:

$$Cov_n = \mathbb{E}\big(x_n y_n\big) \tag{26.3}$$

The correlation estimate on day n is:

$$\rho = \frac{Cov_n}{\sqrt{Var_{x,n} Var_{y,n}}} \tag{26.4}$$

26.4 EWMA

The formula for updating a covariance estimate in the EWMA model is:

$$Cov_n = \lambda Cov_{n-1} + \big(1-\lambda\big) x_{n-1} y_{n-1} \tag{26.5}$$

The lower the value of λ, the greater the weight that is given to recent observations.

26.5 GARCH

GARCH(1, 1) model for updating a covariance rate between X and Y is:

$$Cov_n = \omega + \alpha_{x_{n-1}y_{n-1}} + \beta Cov_{n-1} \tag{26.6}$$

The long-term average covariance rate is:

$$Cov_L = \frac{\omega}{1-\alpha-\beta} \tag{26.7}$$

26.6 Consistency Condition For Covariances

To ensure that a positive-semi-definite matrix is produced, variances and covariances should be calculated consistently.

For example, if variance rates are calculated by giving equal weight to the last m data items, the same should be done for covariance rates. If variance rates are updated using EWMA model with a given λ, the same should be done for covariance rates.

Using a GARCH model to update a variance-covariance matrix in a consistent way requires a multivariate GARCH model.

Chapter 27

Multivariate Normal Distributions

27.1 Properties

Multivariate normal distributions are well understood and easy to deal with. As we see later, they can be useful tools for specifying the correlation structure between variables, even when the distributions are not normal.

Let us consider a bivariate normal distribution of V_1 and V_2. Suppose, that we know the value V_1. Conditional on this, the value of V_2 is normal with mean:

$$\mu = \mu_2 + \rho\sigma_2 \frac{V_1 - \mu_1}{\sigma_1} \tag{27.1}$$

and standard deviation

$$\sigma = \sigma_2 \sqrt{1 - \rho^2} \tag{27.2}$$

Note that

- the expected value of V_2 conditional on V_1 is linearly dependent on the value of V_1
- the standard deviation of V_2 conditional on the value of V_1 is the same for all values of V_1

27.2 Generating Random Samples from Normal Distributions

When samples ϵ_1 and ϵ_2 are required, the usual procedure involves:

1. obtaining independent samples Z_1 and Z_2 from a univariate standard normal distribution
2. the required samples ϵ_1 and ϵ_2 are then calculated as follows:

$$\epsilon_1 = Z_1 \tag{27.3}$$

$$\epsilon_2 = \rho Z_1 + Z_2\sqrt{1-\rho^2} \tag{27.4}$$

If we require samples from a multivariate normal distribution (where all variables have mean zero and standard deviation of one) and the correlation coefficient is $\rho_{i,j}$, we

- first sample n independent variables Z_i where $1 \leq i \leq n$ from a univariate standard normal distributions
- construct samples as:

$$\epsilon_i = \sum_{k=1}^{i} \alpha_{j,k} Z_k \tag{27.5}$$

We must have:

$$\sum_{k=1}^{i} \alpha_{j,k}^2 = 1, \qquad \text{for } 1 \leq j \leq i \tag{27.6}$$

and

$$\sum_{k=1}^{i} \alpha_{i,k}\alpha_{j,k} = \rho_{i,j}, \qquad \forall j < i \tag{27.7}$$

27.3 Factor Models

Sometimes the correlations between normally distributed variables are defined using a factor model.

Suppose that $U_1, U_2, ..., U_N$ have standard normal distributions.

In a one-factor model, each U_1 has a component dependent on a common factor, F, and a component that is uncorrelated with the other variables. Formally:

$$U_i = a_i F + \sqrt{1 - a_i^2} Z_i \tag{27.8}$$

where:

- $F \sim N(0,1)$ and $Z_i \sim N(0,1)$
- and a_i is a constant between -1 and $+1$

The Z_i are uncorrelated with each other and uncorrelated with F.

The coefficient of Z_i is chosen so that U_i has a mean of zero and a variance of one.

In this model the correlation between U_i and U_j arises from their dependence on the common factor F. The correlation coefficient between U_i and U_j is $a_i a_j$.

A one-factor model imposes some structure on the correlations and has the

advantage that the resulting covariance matrix is always positive-semi-definite. Without assuming a factor model, the number of correlations that have to be estimated for the N variables is $N(N-1)/2$. With the one-factor model we need to estimate only N parameters $a_1, ..., a_n$.

Chapter 28

Copulas

28.1 Marginal Distributions

If V_1 and V_2 are two correlated random variables. The **marginal distribution** of V_1 (sometimes also referred as the unconditional distribution) is its distribution assuming we know nothing about V_2.

If the marginal distributions of V_1 and V_2 are normal, a convenient and easy-to-work-with assumption is that the joint distribution of the variables is bivariate normal.

But often there is no natural way to define of defining a correlation structure between two marginal distributions.

28.2 Gaussian Copula

To apply a Gaussian copula for variables V_1 and V_2, the following is done:

- map V_1 and V_2 into new variables U_1 and U_2 that have standard normal

distribution. The mapping is accomplished on a percentile-to-percentile basis. The one-percentile point of the V_1 distribution is mapped to the one-percentile point of the U_1 distribution; the 10-percentile point of the V_1 distribution is mapped to the 10-percentile point of the U_1 distribution; and so on. V_2 is mapped into U_2 in similar way.

- the variables U_1 and U_2 have normal distribution. We assume that they are jointly bivariate normal.
- this is in tern implies a joint distribution and a correlation structure between V_1 and V_2

The essence of copula is therefore that, instead of defining a correlation structure between V_1 and V_2 directly, we do so indirectly. We map V_1 and V_2 into other variables which have well-behaved distributions and for which it is easy to define a correlation structure.

The correlation between U_1 and U_2 is referred as copula correlation. This is not, in general, the same as the correlation coefficient between V_1 and V_2. Because U_1 and U_2 are bivariate normal, the conditional mean of U_2 is linearly dependent on U_1 and the conditional standard deviation of U_2 is constant. However, a similar result does not in general apply to V_1 and V_2.

28.2.1 Expressing the approach algebraically

Suppose, that G_1 and G_2 are the cumulative marginal (i.e. unconditional) probability distributions of V_1 and V_2.

We map so that:

$$G_1(v_1) = N(u_1) \tag{28.1}$$

$$G_2(v_2) = N(u_2) \tag{28.2}$$

where N is the cumulative standard normal distribution.

This means that:

$$u_1 = N^{-1}\big[G_1(v_1)\big] \tag{28.3}$$

$$u_2 = N^{-1}\big[G_2(v_2)\big] \tag{28.4}$$

$$v_1 = G_1^{-1}\big[N(u_1)\big] \tag{28.5}$$

$$v_2 = G_2^{-1}\big[N(u_2)\big] \tag{28.6}$$

The variables U_1 and U_2 are then assumed to be bivariate normal.

The key property of a copula model is that id preserves the marginal distributions of V_1 and V_2 (however unusual these may be), while defining a correlation structure between then.

28.3 Student t-copula

To sample from a bivariate Student's t-distribution with f degrees of freedom and correlation ρ, the steps are as follows:

- sample from the inverse chi-squared distribution to get a value χ
- sample from a bivariate normal distribution with correlation ρ as described earlier
- multiply the normally distributed samples by $\sqrt{f/X}$

It is more common for the two variables to have tail values at the same time in the bivariate Student's t-distribution than in the bivariate normal distribution. To put this another way, the tail dependence is higher in a bivariate Student's t-distribution than in a bivariate normal distribution.
This is called **tail dependence**.

28.4 Definition

A copula is a statistical tool, that creates a joint probability distribution, which assesses the dependence between the variables by retaining their marginal distributions.
A copula is a multivariate probability distribution for which the marginal probability distribution of each variable is uniform.
Copulas are used to describe the dependence between random variables.
Slar's theorem states that any multivariate joint distribution can be written in terms of univariate marginal distribution functions and a copula which describes the dependence structure between the variables.

Chapter 29

Monte Carlo Simulations

29.1 Ways of Choosing a Probability Distribution for a Simulation Model

The ways are:

1. bootstrapping technique
2. parameter estimating technique
3. best fit technique
4. the subjective guess technique

29.2 Motivations

In econometrics, simulation is particularly useful when models are very complex or sample sizes are small.

29.3 Steps

The steps:

1. specify the model that will be used to generate the data

2. estimate the parameter of the interest in the study

29.4 Variance Reduction Techniques

The sampling variation in a Monte-Carlo study is measured by the standard error estimate, denoted S_x:

$$S_x = \sqrt{\frac{var(x)}{N}} \tag{29.1}$$

where N is the number of samples and $var(x)$ is the variance of the estimates of the quantity of interest.

29.4.1 Antithetic Variance

The antithetic variate techniques involves taking the complement of a set of random numbers and running a parallel simulation on those. The variance of the estimations will be reduced because of the negative covariance.
It may at first appear that the reduction in Monte Carlo sampling variation from using antithetic variates will be huge, as the covariance appears to be -1. However, it is important to remember that the relative covariance is between the simulated quantity of interest for the standard replications and those using the

antithetic variates. But the perfect negative covariance is between the random draws and their antithetic variates.

29.4.2 Quasi Monte Carlo

The use of low-discrepancy sequences leads the Monte-Carlo standard errors to be reduced in direct proportion to the number of replications rather than in proportion to the square root of the number of replications.

29.4.3 Control Variates

The application of control variates involves employing a variable similar to that used in simulation, but whose properties are known prior to the simulation.
It is worth noting that control variates succeed in reducing the Monte Carlo sampling error only if the control and simulation problems are very closely related.

29.4.4 Random Number Re-Usage across Experiments

29.4.5 Bootstrapping

Bootstrapping is used to obtain a description of the properties of empirical estimators by using the sample data points themselves, and it involves sampling repeatedly with replacement from the actual data.
The advantage of bootstrapping over the use of analytical results is that it allows the researcher to make inferences without making strong distributional assumptions, since the distribution employed will be that of the actual data.

Instead of imposing a shape on the sampling distribution of the θ value, bootstrapping involves empirically estimating the sampling distribution by looking at the variance of the statistic within-sample.
Successive applications of this procedure should generate a collection of data sets with the same distributional properties, on average, as the original data. But any kind of dependence in the original series (e.g. linear on non-linear autocorrelation) will, by definition, have been removed.

Situations Where Bootstrapping Will Be Ineffective

There are at least two situations where the bootstrap will not work well:

- outliers in the data. If there are outliers in the data, the conclusions of the bootstrap may be affected. In particular, the results for a given replication may depend critically on whether the outliers appear (and how often) in the bootstrapped sample.
- non-independent data. **Bootstrapping implicitly assumes that the data are independent of one another**. This would obviously not hold if, for example, there were autocorrelation in the data.

29.5 Latin Hypercube

A square grid containing sample positions is a Latin square if (and only if) there is only sample in each row and each column. A Latin hypercube is the generalisation of this concept to an arbitrary number of dimensions, whereby each sample is the only one in each axis-aligned hyperplane containing it.
When sampling a function of N variables, the range of each variable is divided into M equally probable intervals. M sample points are then placed to satisfy

the Latin hypercube requirements. Note, that this forces the number of divisions M to be equal for each variable.

Also note, that this sampling scheme does not require more samples for more dimensions (variables). This independence is one of the main advantages of this sampling scheme.

Another advantage is that random samples can be taken one at a time, remembering which samples were taken so far.

In two dimensions the difference between random sampling, Latin Hypercube sampling and orthogonal sampling can be explained as follows:

- **random or brute-force sampling** - new sample points are generated without taking into account the previously generated sample points. One does not necessarily need to know beforehand how many sampling points are needed.
- in **Latin Hypercube sampling** one must first decide how many sample points to use and for each point remember in which row and column the sample point was taken.
- in **Orthogonal sampling**, the sample space is divided into equally probable subspaces. All sample points are hen chosen simultaneously making sure that the total ensemble of sample points is a Latin Hypercube sample and that each subspace is sampled with the same density.

Thus:

- **orthogonal sampling** - ensures that the ensemble of random numbers is a very good representative of the real variability
- **Latin Hypercube sampling** - ensures that the ensemble of rnadom numbers is representative of the real variability

- **traditional random sampling** - is just an ensemble of random numbers without any guarantees.

29.6 Random Number Generation

29.6.1 Pseudo-random numbers

Numbers that are a continuous uniform (0, 1) can be generated according to the following recursion:

$$y_{i+1} = \left(a \cdot y_i + c\right) \quad \text{modulo m}, i = 0, ..., T \tag{29.2}$$

Then:

$$R_{i+1} = y_{i+1}/m \quad \text{for } i = 0, 1, ..., T \tag{29.3}$$

for T random draws, where:

- y_0 is the seed (the initial value of y),
- a is a multiplier
- and c is an increment

All three of these are simply constants. The "modulo operator" simply functions as a clock, returning to one after reaching m.
Any simulation study involving a recursion, such as that described by the equation above, to generate the random draws, will require the user to specify an initial value y_0 to get the process started. The choice of this value will, undesirably, affect the properties of the generated series. This effect will be strongest

for $y_1, y_2, ...$ but will gradually die away. Consequently, a good simulation design will allow for this phenomenon by generating more data than required and then dropping the first few observations.

29.6.2 Mid-square Technique

Steps:

- a 4-digit starting value is created and squared, producing an 8-digit number.
- if the result is fewer than 8 digits, leading zeros are added to compensate
- the middle 4 digits of the result would be the next number in the sequence and returned as the result
- the process is repeated to generate more numbers

29.7 Disadvantages of the Simulation Approach to Econometric or Financial Problem Solving

- it might be computationally expensive
- the results might not be precise, for example, if some unrealistic assumptions have been made of the data generating process.
- the results are often hard to replicate
- simulations results are experiment specific

29.8 Monte Carlo Biases

- **discretisation bias** - when a variable is incorrectly assumed to take on only discrete values