ECOM40006/ECOM90013 Econometrics 3
Department of Economics
University of Melbourne

An Introduction to Large-Sample Asymptotics

Semester 1, 2025

Version: March 19, 2025

Treatments of asymptotic arguments frequently come in one of two forms, either (i) careful but technically advanced and somewhat inaccessible for those with limited preparation, or (ii) accessible but technically loose and sometimes erroneous. The purpose of this set of notes is to provide a treatment of large-sample asymptotics that is careful yet not so advanced as to be inaccessible. It is meant to be something of a half-way house to these more advanced treatments. Moreover, it is hoped that the treatment is sufficiently self-contained that those with limited preparation can cope with the material.

One final warning, the material provided in the appendices is there for those who want to dive right in. But it is in the appendices because it is over and above what we are doing in this subject. So, if you find it all a bit daunting then just ignore it.

Contents

1 Introduction

Asymptotic arguments, leading to approximations for exact sampling behaviour, have proved to be the cornerstone of much statistical analysis. The general idea is the following:

> Suppose that you have a problem which can be characterized in terms of some index $n \in \mathbb{N}$, where $\mathbb{N}$ denotes the set of *natural numbers* or positive integers. We shall write P_n to denote our problem. You can think of nesting your problem in an ordered sequence of such problems, *viz.*
>
> $$P_1, P_2, \ldots, P_{n-1}, P_n, P_{n+1}, \ldots$$
>
> The purpose of the index is to indicate order in the sequence, the larger the value of the index the deeper in the sequence you are. Notionally any one of these problems may be too difficult to solve directly but we may be able to say something about the limiting case, denoted P, which is that problem that the sequence approaches as we allow the index to become infinitely large. One might then argue that if the sequence converges monotonically towards P, and if n is sufficiently large, so that P_n is sufficiently deep in the sequence, then the solution for P may be very similar to that for P_n and so might provide a good approximation to the solution to P_n.

In summary, the fundamental idea of asymptotics is that our problem is too hard to solve directly and so we approximate the solution to this difficult problem using the solution to a problem that we can solve and which might, under certain conditions, be expected to be close to the solution we seek. The fact that such an approximation might only be expected to work well *under certain conditions* should signal to you that there are times when so-called asymptotic approximations work very well and times when they perform very poorly. Typically approximations work poorly when the difference(s) between the approximation and the exact solution become important considerations and asymptotic approximations are no different in this regard.

Why are we interested in large-sample approximations? In order to address this question we need to think about the sequence of problems in which we nest, or embed, our problem P_n; for large-sample asymptotics the index denotes sample size. The simple fact of the matter is that this sequence is completely artificial. There are no problems P_{n+1}, P_{n+2}, ..., they are just figments of our imagination, our sample size is n and not $n+1$ or $n+2$, etc.[1] So why do we allow our imagination to wander in this way? It is easy to imagine (time-series) situations where sample sizes grow. For example, suppose that you are working with national accounting data where new observations are released every quarter. This means that every three months your sample size gains an extra observation and the sequence of problems in which you have nested your problem of interest is simply reflecting the passage of time. Of course, you don't want to lean on this intuition too hard because allowing the index to pass to infinity means that you are thinking of an infinite time horizon and that is something that none of us will ever get to see. Furthermore, it is extremely difficult to imagine a world where important features of our problem will remain fixed over an infinite time horizon.

Another way of motivating such a nesting sequence is to think of the limiting case of an infinitely large sample as one where the entire population is observed. So our nesting

[1] Interestingly, if the index denotes sample size then there are problems $P_1, \ldots, P_{n-1}$ but they would involve throwing away sample observations which is typically something that we don't want to do.

sequence is one where the sample becomes increasingly close to a census of the population as the index (sample size) increases. This is probably how most people would describe the role of asymptotics in cross-section, or microeconometric, situations, although it is far from perfect. At the very least you need to ask yourself just how many infinitely large populations you actually encounter.

We see that it is very difficult to provide good intuition as to where this nesting sequence comes from and yet large-sample asymptotics are used a lot. The question is why? We don't actually need a good physical example of the nesting sequence. It is a purely hypothetical mathematical device and, as such, can be chosen completely arbitrarily. Our objectives in choosing a nesting sequence are two-fold. First, in whichever sequence we choose to embed our original problem, it is essential that we can actually solve the limiting form of the problem, namely P. Simply put, no solution means no approximation. Second, if we restrict attention to sequences for which we can solve the limiting problem, the only important criteria is that the limiting solution provide a good approximation to the problem in which we are interested, namely P_n. In the same way as the proof of a pudding is in the eating, the usefulness of a nesting sequence is in the quality of the approximation that it yields. Clearly there is no compelling reason why any sequence need be indexed by the sample size and, indeed, there are other sequences encountered in the literature.[2] Nevertheless, it has been found in many, many situations that nesting the problem of interest in a sequence of problems distinguished one from the other by sample size yields limiting cases that can be solved and whose solution provide extremely good approximations to the solution of the original problem. That is, we use large-sample asymptotics both because we can and because they often seem to yield good approximations.

Our plan of attack on asymptotics shall be the following. First, we will address the notion of convergence. P is that problem to which the sequence $P_1, P_2, \ldots$ converges. Obviously we cannot characterize the convergence of P_1, P_2, ... without knowing what is meant by convergence. Second, having defined a suitable notion of convergence and hence the problem P, it is necessary that that we be able to solve the problem. We shall refer to this solution as a limiting result. Finally, once we have a limiting result we need to know how to use it to approximate the solution to our problem of interest, namely P_n. This approximate solution is then called an asymptotic approximation. We shall provide a number of examples along the way but our primary focus is on large-sample asymptotics as they apply to maximum likelihood methods.

The treatment of asymptotic arguments provided here is not meant to be definitive. More advanced (and complete) treatments can be found in references such as Amemiya (1985), Cramér (1946), Gallant (1997), Newey and McFadden (1994), Rao (1973), Schmidt (1976), Theil (1971), and Wald (1949). This list is far from exhaustive.

2 Notions of Convergence

Consider the limit of the sequence of numbers $X_1, X_2, \ldots$, where the typical term in the sequence is $X_n = 1 + n^{-1}$. It is clear that as we allow n to become infinitely large this sequence of numbers will converge to unity. By 'converge' here we mean that by allowing n to become sufficiently large we can ensure that each subsequent term in the sequence

[2] Typically these other sequences differ only in the interpretation of the index set. The notions of convergence, etc. remain unchanged.

lies within some interval centred at unity, where the width of the interval can be chosen a priori to be any positive number as small as we like. It is interesting to note that there is no finite value of n for which $X_n = 1$ and so it is important to recognize from the beginning that it does not follow that the limit of a sequence is necessarily a point in the sequence.

We know all about the sequence $X_n = 1+n^{-1}$, $n = 1, 2, \ldots$, with certainty because the sequence is comprised of (non-random) numbers. When we have to deal with sequences of random variables the entire notion of convergence becomes more complicated because the randomness of these variables means that we can't know with certainty exactly what values they will take. What can be known about random variables is that they possess cumulative distribution functions. Consequently our notions of convergence will be cast in terms of these functions and their properties.

We shall mention four notions of convergence: convergence in distribution, convergence in probability, convergence in rth mean, and almost sure convergence. These different modes of convergence differ in what they assume/require of the problem. The strongest results have the most stringent requirements. For our purposes the weaker results will be sufficient and they will be where we focus our attention.

2.1 Convergence in Distribution

This is the weakest form of convergence that we will consider. It says something about the probabilities of events of the form $X_n \leq x$ and $X \leq x$ but it says nothing about the proximity of the random variables X_n and X. Let us begin with a definition and then explore some examples.

Convergence in Distribution.
A sequence $X_1, X_2, \ldots$ is said to converge to X in distribution if the distribution functions F_n of X_n converge to the distribution function F of X at every continuity point of F. That is,

$$\lim_{n\to\infty} |F_n(x) - F(x)| = 0.$$

We write $X_n \xrightarrow{d} X$ and we call F the limiting distribution of the sequence $X_1, X_2, \ldots$

This is our first chance to think about what is meant by the proximity of random variables. To begin, consider two Bernoulli random variables Y and W with same probability of taking the value unity, $\Pr(Y = 1) = \Pr(W = 1)$. Just because they have the same probability of taking the value unity does not imply that Y will always take the value 1 when W does if draws are taken randomly from each of these two populations.

We might extend this simple example to consider two random variables Z_1 and Z_2, each with standard normal distributions. Clearly $\Pr(Z_1 \leq z) = \Pr(Z_2 \leq z)$ for all $z \in \mathbb{R}$, where $\mathbb{R}$ denotes the set of real numbers. However, if we were to take draws from each of these distributions we might, for example, observe $Z_1 = -27,306$ and $Z_2 = 7$. These observed values are very different from one another. Moreover, suppose now that $z = 0$. We see here that the event of being less than or equal to $z = 0$ is only satisfied by one of the two random variables, Z_1, even though both Z_1 and Z_2 have the same probability of satisfying the event.

Now consider the sequence of random variables $X_1, X_2, \ldots$ and suppose that for some constant x, say $x = 10$, $\Pr(X_n \leq 10) = 0.2 + (1 + n)^{-1}$. As $n \to \infty$ it is clear that $\Pr(X_n \leq 10) \to 0.2$. This is the limiting probability of this event. If we can find a

random variable X such that $\Pr(X_n \leq x) \to \Pr(X \leq x)$ for all possible choices of x then we say that the distribution of X is the limiting distribution of the sequence $X_1, X_2, \ldots$; that is, $X_1, X_2, \ldots$ converges in distribution to X.

There are many different symbols that people use to denote convergence in distribution. For example, sometimes it is denoted by a capital D, or perhaps $\mathcal{D}$, e.g. $\xrightarrow{D}$ or $\xrightarrow{\mathcal{D}}$. This type of convergence is also sometimes called convergence in law, which reflects the fact that probability distributions used to be called probability laws. This leads to notation such as $X_n \xrightarrow{L} X$ or $X_n \xrightarrow{\mathcal{L}} X$. I have also seen notation such as $X_n \equiv X$ used to denote the same idea. So it is always important that you pay close attention to the use of notation when thinking about these ideas.

2.2 Convergence in Probability

Convergence in distribution only tells us about the probabilities of random variables being observed in regions of their support, without telling us anything about where the random variable will actually be. Oftentimes we want to say something stronger. Convergence in probability, also known as weak convergence, goes some way towards allowing us to make such statements.

Convergence in Probability.
A sequence of random variables $X_1, X_2, \ldots$ is said to converge in probability to a random variable X if

$$\lim_{n\to\infty} \Pr(|X_n - X| \geq \epsilon) = 0, \quad \text{for any } \epsilon > 0.$$

If this condition is satisfied we write $X_n \xrightarrow{p} X$.

We may alternatively define convergence in probability according to

$$\lim_{n\to\infty} \Pr(|X_n - X| < \epsilon) = 1, \quad \text{for any } \epsilon > 0.$$

These definitions are completely equivalent and both are encountered in the literature.

Rather than writing $X_n \xrightarrow{p} X$ another commonly encountered notation for convergence in probability is

$$\operatorname*{plim}_{n\to\infty} X_n = X,$$

or simply $\operatorname{plim} X_n = X$ if the sequence is clear from the context, where plim is short for 'probability limit'.

In order to read the probability statements that define convergence in probability start in the middle and work out. We will work with the first definition but you should ensure that you understand the logic of the second definition as well. First, $|X_n - X| \geq \epsilon$ defines the event that the distance between X_n and X is greater than or equal to ϵ. Note that the use of the absolute value function $|\cdot|$ means that we don't care whether X_n is bigger or smaller than X, we only care about how far apart they are.[3] Also note that ϵ

[3]The absolute value function is defined as

$$|x| = \begin{cases} x, & x \geq 0, \\ -x, & x < 0. \end{cases}$$

is defined to be any positive real number. In particular, we can choose ϵ to be as small as we wish, the technical expression for this is 'arbitrarily close to zero'.

Next we have $\Pr(|X_n - X| \geq \epsilon)$. That is, convergence in probability is concerned with the probability of X_n and X being far apart, where 'far apart' means any distance greater than or equal to ϵ and ϵ, as you should recall, can be as small a positive number as you wish. Convergence in probability therefore tells us that if we consider the sequence of probabilities of the events $|X_n - X| \geq \epsilon$, as $n \to \infty$, then we see that these probabilities tend to zero. That is, as n becomes large the probability of there being any distance at all between X_n and X becomes vanishingly small.

Now we need to be careful when thinking about what convergence in probability actually means. The probability of anybody winning tattslotto is essentially zero and yet somebody, and often more than one ticket, wins it almost every week. Similarly, the probability of the two random variables X_n and X taking different values is essentially zero for sufficiently large n *but they can still differ* because it is only the *probability* of them differing which is converging to zero and not the differences between the realizations of the random variables themselves.

The simplest example of convergence in probability is when the random variable of interest converges to a constant. In order for a random variable to converge to a constant it must be the case that as $n \to \infty$ the distribution of X_n is collapsing about that constant, essentially becoming a spike which is located at that point. As an example, suppose that we have a Bernoulli random variable Y_n. Now suppose that we define a new random variable according to $Z_n = C + Y_n/n$, for some constant C, so that

$$Z_n = \begin{cases} C, & \text{if } Y_n = 0, \\ C + n^{-1}, & \text{otherwise.} \end{cases}$$

We see then that

$$\lim_{n \to \infty} \Pr(|Z_n - C| \geq \epsilon) = 0.$$

That is, $\operatorname{plim} Z_n = C$. This final result holds because no matter how small is the ϵ that we choose there will always exist a value n^* such that $n^{-1} \leq \epsilon$ for all $n > n^*$.

As a further example, suppose that in the previous example C had been a random variable rather than a constant. We still have the result $Z_n \xrightarrow{p} C$. Here we see that it is the distribution of difference between Z_n and the random variable that is its probability limit which is collapsing to a spike at zero.

It is interesting to note that that convergence in probability must also imply convergence in distribution. Given that the probability of Z_n being different from C tends to zero as $n \to \infty$, it follows that the probabilities of events of the form $Z_n \leq c$ and $C \leq c$ must also be the same in the limit. Therefore, we also have $Z_n \xrightarrow{d} C$. If C is a constant then Z_n has a degenerate limiting distribution whereas if C is random then Z_n has the distribution of C as its limiting distribution.

2.3 Convergence in rth Mean

Convergence in probability is often the sort of property that we are seeking to establish but it can sometimes be hard to prove directly. Luckily it is implied by convergence in rth mean which is often easier to establish. Unfortunately this ease of use comes at a cost in terms of generality, although the cost may be small and one that you are willing to pay. Before exploring this trade-off let us define the mode of convergence.

Convergence in rth Mean.
A sequence $X_1, X_2, \ldots$ is said to converge to X in rth mean if $\mathrm{E}\left[\left|X_n\right|^r\right] < \infty$ and $\mathrm{E}\left[|X|^r\right] < \infty$, and

$$\lim_{n\to\infty} \mathrm{E}\left[|X_n - X|^r\right] = 0.$$

We write $X_n \xrightarrow{r} X$.

Convergence in rth mean is also called L_r convergence and we would write X_n converges in L_r to X.

This criterion imposes stronger conditions on the problem than does convergence in probability because it requires the existence of absolute moments for all n. If you think of an expectation as a probability weighted sum, and take into account that the absolute value function here means that all of the terms being weighted together are non-negative, this condition requires that tail probabilities be sufficiently small that the expectation is finite. Convergence in probability makes no assumptions about the existence of moments and, in particular, places no such requirements on the magnitudes of tail probabilities. That said, in many circumstances it is not a strong assumption for the relevant moments to exist and so there is no reason not to establish convergence in pth mean. As we shall see, the special case of convergence in mean square ($r = 2$) is frequently used to establish convergence in probability.

To see that convergence in rth mean implies convergence in probability we need to start with Markov's inequality:

$$\Pr\left(g(Z) \geq \epsilon\right) \leq \frac{\mathrm{E}\left[g(Z)\right]}{\epsilon},$$

where $g(Z)$ is a non-negative function of the random variable Z and $\epsilon > 0$. If we then set $g(X_n) = |X_n - X|$ then we obtain

$$\Pr\left(|X_n - X| \geq \epsilon\right) \leq \frac{\mathrm{E}\left[|X_n - X|\right]}{\epsilon}. \tag{1}$$

Now suppose that X_n converges in rth mean to X. If an expectation exists for any value of r it must also exist for smaller values of r so, for example, a variance can only exist if the mean also exists. The consequence of this observation is that if convergence in rth mean occurs at all it must hold for $r = 1$. If we now take limits of both sides of (1) as $n \to \infty$ we obtain

$$\lim_{n\to\infty} \Pr\left(|X_n - X| \geq \epsilon\right) \leq \lim_{n\to\infty} \frac{\mathrm{E}\left[|X_n - X|\right]}{\epsilon} = \frac{0}{\epsilon} = 0,$$

where the first equality follows from the supposition of convergence and the second equality follows because $\epsilon > 0$. That is, $\lim_{n\to\infty} \Pr\left(|X_n - X| \geq \epsilon\right) \leq 0$. But probabilities cannot be negative which implies that $\lim_{n\to\infty} \Pr\left(|X_n - X| \geq \epsilon\right) = 0$ as required.

In general, working with mean absolute convergence ($p = 1$) is much the same as working with the original definition of convergence in probability. Two results that are sometimes helpful in this context are Hölder's inequality and Minkowski's inequality, which are briefly discussed in Appendix B.

Typically it is easier to work with $r = 2$, in which case we see that

$$|X_n - X|^2 = (X_n - X)^2.$$

Here we end up working with mean squared errors, which are things that we should be used to seeing. However, there is a cost in that we require variances to exist which, again, is a restriction on tail probabilities; if they are too large the variances will not exist because the relevant sums or integrals become infinitely large.[4]

Convergence in Mean Square (or Convergence in Quadratic Mean).
A sequence $X_1, X_2, \ldots$ is said to converge in mean square to X, written $X_n \xrightarrow{m} X$, if

$$\lim_{n\to\infty} \mathrm{E}\left[(X_n - X)^2\right] = 0.$$

Convergence in mean square is often a relatively simple way of establishing convergence in probability to a constant. Observe that, for any constant c,

$$\begin{aligned}\mathrm{E}\left[(X_n - c)^2\right] &= \mathrm{E}\left[(X_n - \mathrm{E}[X_n])^2\right] + (\mathrm{E}[X_n - c])^2 \\ &= \mathrm{Var}[X_n] + (\mathrm{E}[X_n - c])^2\end{aligned}$$

which is the usual decomposition of mean squared error into variance and bias squared. If it can be established that $\lim_{n\to\infty} \mathrm{Var}[X_n] = 0$ and $\lim_{n\to\infty} \mathrm{E}[X_n] = c$ then it follows that $X_n \xrightarrow{m} c$ which, by our earlier result, implies that $X_n \xrightarrow{p} c$. This is a very common approach in large-sample asymptotics.

2.4 Almost Sure Convergence

Almost sure convergence, also known as strong convergence, is a stronger form of convergence than is convergence in probability. When applying our convergence results we will restrict attention to convergence in probability although many of our results, if not all, can be established almost surely.

Almost Sure Convergence.
A sequence of random variables $X_1, X_2, \ldots$ is said to converge almost surely to a random variable X if

$$\Pr\left(\lim_{n\to\infty} |X_n - X| \geq \epsilon\right) = 0, \quad \text{for any } \epsilon > 0.$$

If this condition is satisfied we write $X_n \xrightarrow{a.s.} X$ or $\lim_{n\to\infty} X_n = X$ a.s.

This form of convergence is also known as *convergence almost everywhere* and *convergence with probability 1* and so you also see the following notational conventions adopted throughout the literature: $\lim_{n\to\infty} X_n = X$ a.e. or $X_n \xrightarrow{a.e.} X$, $\lim_{n\to\infty} X_n = X$ wp1 or $X_n \xrightarrow{wp1} X$.

The distinction between almost sure convergence and convergence in probability is a subtle one and to understand it we need go back to the fundamental definition of probability and define what we mean by a *set of measure zero*. Let us think of some random experiment as possessing a sample space of possible distinct outcomes. We shall partition this sample space into a set defining some event, $\mathcal{A}$ say, and the complement of this set, which contains those outcomes that correspond to the event $\mathcal{A}$ not occurring (the complementary event, $\overline{\mathcal{A}}$). If the sample space contains a finite number of equally likely outcomes we might define the probability of $\mathcal{A}$ occurring as the proportion of these

[4] By implication, the larger the value of r that one attempts to work with the more stringent are the requirements on tail probabilities and so the less general the result. It would be very unusual to work with $r > 2$. Remember that convergence in probability does not require that any moments exist.

elementary outcomes contained in the set defining $\mathcal{A}$. Now suppose that we allow the number of outcomes in the sample space to become infinitely large but we hold fixed the outcomes in the set defining $\overline{\mathcal{A}}$. Clearly $\Pr(\mathcal{A})$ approaches unity as our set of possible outcomes becomes infinitely large. However, there remains elements in the set $\overline{\mathcal{A}}$ and so it is not empty, it is just that the probability ascribed to this set, which has become an infinitesimally small proportion of the sample space, is zero. We speak of the event $\overline{\mathcal{A}}$ as being a set of measure zero. You have actually encountered such sets many times before without them necessarily being described in this way. For example, consider a (continuous) normally distributed random variable, Z say. Suppose that we take a draw from the population of Z and observe some number c. Then the event $Z = c$ has occurred even though we know that $\Pr(Z = c) = 0$ for any continuous random variable. Simply put c is a set of measure zero, i.e. the probability of observing c is zero but the set comprised of c is not empty. For all continuous random variables, with non-degenerate distributions, single values are sets of measure zero.

Both convergence in probability and almost sure convergence can be cast in terms of exceptional sets of measure zero, with the exceptional sets comprised of those values of X_n where convergence fails, i.e. where $|X_n - X| \geq \epsilon$. The difference between the two notions of convergence is how these exceptional sets are defined. Almost sure convergence looks at the limiting case of this set, that is $\lim_{n\to\infty} |X_n - X| \geq \epsilon$. This is a once and for all definition of the exceptional set and the probability ascribed to elements of that set is always zero. In contrast, convergence in probability considers a sequence of such sets and membership of the exceptional set may therefore vary with the index n. Moreover, it is only in the limit that this set is a set of measure zero. In particular, for any finite n, there can be a non-zero probability of divergence between X_n and X. So, convergence in probability is the weaker concept, but it is easier to establish and that is why it is so important.

Note that, except for certain special cases, almost sure convergence does not imply convergence in rth mean nor does convergence in rth mean imply almost sure convergence, although both imply convergence in probability. We will tend not to work with almost sure convergence for the simple reason that strong convergence results can be very difficult to establish.

2.5 Useful Convergence Results

Our first result is due to Slutsky.[5]

Theorem 1 (Slutsky's Theorem)**.** *If* $X_n \xrightarrow{d} X$ *and* $Y_n \xrightarrow{p} c$*, where* c *is a constant, then*

(i) $X_n + Y_n \xrightarrow{d} X + c$

(ii) $X_n Y_n \xrightarrow{d} cX$

(iii) $X_n / Y_n \xrightarrow{d} c^{-1}X$*, provided that* $c \neq 0$. □

One thing to take out of these results is that they do not simply present a convenient way that you might find the relevant limits. Rather they indicate the place of limits in

[5] Evgeny Evgenievich Slutsky (1880–1948) made important contributions to both probability theory and also microeconomics. According to the relevant Wikipedia page https://en.wikipedia.org/wiki/Slutsky%27s_theorem, Slutsky's Theorem is also attributed by some to Harald Cramér.

the order of operations. For example, an important application of Theorem 1(ii) comes from the observation that $X_nY_n = X_nc + X_n(Y_n - c)$. The term $Y_n - c \xrightarrow{p} 0$ and so, by Theorem 1(ii), the expression $X_n(Y_n - c) \xrightarrow{d} 0$. That is, this expression has a degenerate distribution located at zero. The implication of this argument is that X_nY_n and X_nc will have the same limiting distribution. This suggests a step-wise approach to obtaining limiting distributions where the first step is to replace terms with probability limits by those limits and the second step then concentrates on finding the limiting distributions of the remaining terms. This is a very commonly encountered approach to the derivation of limiting distributions.

Note that if, in Theorem 1, $X_n \xrightarrow{p} X$, where now X is a constant, then the results of the corollary will still hold except that the mode of convergence will be convergence in probability rather than convergence in distribution. For example, $X_n + Y_n \xrightarrow{p} X + c$.

It is sometimes useful to recognize that if Y_n has a limiting distribution and $\operatorname{plim}(X_n - Y_n) = 0$ then X_n has the same limiting distribution as Y_n. The argument behind this result was given at the end of Section 2.2.

We have the following important extensions of these results to matrix random variables, which obviously includes vectors as a special case. Note that when we talk of the probability limit of a matrix we are thinking of the limits of the individual elements. That is, for an $p \times q$ matrix W_n,

$$\operatorname{plim} W_n = \{\operatorname{plim}(W_n)_{s,t}\}_{\substack{s=1,\ldots,p \\ t=1,\ldots,q}}.$$

In the results that follow we will assume that the stated matrices are conformable as required.

1. If $\operatorname{plim} X_n = A$ and $\operatorname{plim} Y_n = B$ then $\operatorname{plim} X_nY_n = AB$.

2. Let g be a real matrix-valued function that is not a function of n and which is continuous at A. If the matrices $X_n \xrightarrow{p} A$ then $g(X_n) \xrightarrow{p} g(A)$. This is a matrix version of Slutsky's theorem.

3. Let g be a real matrix-valued function that is not a function of n and which is continuous at X. If the matrices $X_n \xrightarrow{d} X$ then $g(X_n) \xrightarrow{d} g(X)$. If the matrix version of Slutsky's theorem applies then this result must follow because convergence in probability implies convergence in distribution.

Next we have an important result that allows us to deal with functions of random variables:

Theorem 2 (Continuous Mapping Theorem)**.** *For any continuous function g that is not a function of n:*

$$\operatorname*{plim}_{n\to\infty} g(X_n) = g\left(\operatorname*{plim}_{n\to\infty} X_n\right). \qquad \square$$

The requirement that g not be a function of n precludes, for example, functions such as $g(X_n) = nX_n^2$.

The importance of this theorem is that it allows us to establish convergence of random variables in the form that is easiest to work with and then map that result into the form of the random variable that we actually need to work with, provided that the mapping is continuous (which it typically will be). It is also worth contrasting this property of

probability limits with its analogue for expectations: $\mathrm{E}\left[g(X_n)\right] \neq g(\mathrm{E}\left[X_n\right])$ unless g is a linear function.

As a final observation, recall that that convergence in probability implies convergence in distribution and so Theorem 2 implies a similar statement about the limiting distributions of continuous functions of random variables

2.6 An Aside on Order

In this section we introduce a new form of notation, the purpose of which is to provide a mechanism whereby we can discuss the rate at which convergence occurs. To begin let us think about sequences of non-random real numbers.

Definition 1. Big O and Little o Notation.
Consider two sequences of real numbers $X_1, X_2, \ldots$ and $Y_1, Y_2, \ldots$ The sequence $X_1, X_2, \ldots$ is said to be at most of order Y_n if

$$\lim_{n\to\infty} \left| \frac{X_n}{Y_n} \right| = c, \quad \text{for some positive constant } c < \infty.$$

We write $X_n = O(Y_n)$. The sequence $X_1, X_2, \ldots$ is said to be of smaller order than Y_n, denoted $X_n = o(Y_n)$, if

$$\lim_{n\to\infty} \left(\frac{X_n}{Y_n} \right) = 0. \qquad \square$$

Let us illustrate these ideas with some examples.

1. Consider $X_n = 6 + n^{-1}$. This sequence of numbers neither diverges nor converges to zero as $n \to \infty$ and so we say that $X_n = O(1)$. With large-sample asymptotics we typically express the orders of variables in terms of powers of the sample size n and so it is also correct (and sometimes helpful) to think of $O(1)$ as $O(n^0)$. Note that if $\lim_{n\to\infty} X_n/n^\delta = 0$, $\delta > 0$, we conclude that $X_n = o(n^\delta)$.

2. Next consider $S_n = \sum_{i=1}^n (6 + n^{-1}) = 6n + 1 = O(n)$. We see that a sum of n terms, each of order $O(1)$, is of order n. If we then calculate an average $\overline{X}_n = n^{-1} S_n = 6 + n^{-1} = O(1)$. Similarly, we see that, $n^\delta S_n = O(n^{1+\delta})$ and $n^\delta \overline{X}_n = O(n^{0+\delta}) = O(n^\delta)$.

3. Let $X_s = 3s^2 - 2$. Clearly, $\lim_{s\to\infty} X_s/s^2 = 3$ and so $X_s = O(s^2)$.

4. Consider next $X_t = (3t^2 - 2)^{-1}$. Here,

$$\lim_{t\to\infty} \frac{X_t}{t^{-2}} = \lim_{t\to\infty} \frac{t^2}{3t^2 - 2} = \lim_{t\to\infty} \frac{1}{3 - 2/t^2} = \frac{1}{3} \implies X_t = O(t^{-2}).$$

5. Suppose that $X_a = (a + 7)/(3a^2 - 2)$. Then

$$\lim_{a\to\infty} X_a = \lim_{a\to\infty} \frac{a + 7}{3a^2 - 2} = \lim_{a\to\infty} \frac{1/a + 7/a^2}{3 - 2/a^2} = \frac{0}{3} = 0.$$

   That is, $X_a = o(1)$. However, if we multiply X_a by a — or, equivalently, divide by a^{-1} — the limit would be $1/3 = O(1)$ from which we infer that $X_a = O(a^{-1})$.

6. If $X_n = \exp(-n)$ then

$$X_n = \frac{1}{1 + n + \frac{1}{2}n^2 + \frac{1}{3!}n^3 + \frac{1}{4!}n^4 + \ldots}$$

Clearly $\lim_{n\to\infty} n^\delta X_n = 0$ for any $\delta > 0$ and so we see $\exp(-n) = o(n^{-\delta})$, $\delta > 0$.

It also is possible to determine the following results.

(i) If $X_n = O(n^\alpha)$ and $Y_n = O\left(n^\beta\right)$, then

$$\begin{aligned} X_n Y_n &= O\left(n^{\alpha+\beta}\right) \\ |X_n|^r &= O\left(n^{\alpha r}\right) \\ X_n + Y_n &= O\left(\max\left\{n^\alpha, n^\beta\right\}\right) \end{aligned}$$

(ii) If $X_n = O(n^\alpha)$ and $Y_n = o\left(n^\beta\right)$, then

$$\begin{aligned} X_n Y_n &= o\left(n^{\alpha+\beta}\right) \\ X_n + Y_n &= O\left(n^\alpha\right) \end{aligned}$$

One important application of these results involves the Taylor expansion which we might write as

$$f(x) = f(a) + f^{(1)}(a)(x-a) + \frac{f^{(2)}(a)}{2!}(x-a)^2 + \ldots + \frac{f^{(n)}(a)}{n!}(x-a)^n + o\left(|x-a|^n\right),$$

as $x - a \to 0$, where $f^{(n)}(a) = [(d^n f(x))/(d^n x)]_{x=a}$. This result provides us with a basis for asymptotic approximation because it allows us to control the order of the remainder by including as many terms from the Taylor expansion as necessary in the approximating function.

We can now extend these ideas to sequences of random variables.

Definition 2. Stochastic Order.
Let $X_1, X_2, \ldots$ be a sequence of random variables and let $Y_1, Y_2, \ldots$ be a sequence of real positive constants. The sequence $X_1, X_2, \ldots$ is said to be at most of order Y_n in probability if there exists a non-stochastic sequence $c_1, c_2, \ldots$ such that $c_n = O(1)$, $n = 1, 2, \ldots$, and $\text{plim}(Y_n^{-1} X_n - c_n) = 0$. We write $X_n = O_p(Y_n)$. If $\text{plim}\, Y_n^{-1} X_n = 0$, written $X_n = o_p(Y_n)$, then X_n is of order smaller than Y_n in probability. □

Note that if the sequence of variables $X_1, X_2, \ldots$ is non-stochastic then we see then this definition yields our earlier definition because

$$\underset{n\to\infty}{\text{plim}}\, X_n = \lim_{n\to\infty} X_n$$

for non-stochastic X_n. In this case there is no need to indicate that you are working with random variables and so the 'Big O-Little o' notation is used without the p subscript.

Earlier we presented results on orders of magnitude of sums and products of sequences of numbers. Similar results exist for sequences of random variables and we will not bother to reproduce them in full. Instead we present some other useful results on orders of magnitude in probability:

1. If $\text{Var}[X_n] = \sigma_n^2 < \infty$ then $X_n = O_p(\sigma_n)$, so the order of magnitude of a random variable is that of its standard deviation.

2. If $X_n = O_p(n^{-1/2})$ then $X_n = o_p(1)$.

3. If $X_n \xrightarrow{d} X$ then $X_n = O_p(1)$. Larger stochastic order for X_n would mean that it is diverging as $n \to \infty$ which is inconsistent with the notion that it is converging to anything. Smaller stochastic order would mean that it has a degenerate limiting distribution, which implies convergence in probability.

4. If $X_n \xrightarrow{d} X$ then $X_n + o_p(1) \xrightarrow{d} X$. We see that adding terms of smaller order to X_n will not impact upon its limiting behaviour.

3 Limiting Results

We consider two types of limiting results, (i) laws of large numbers and (ii) central limit theorems. The primary difference between the two types of results, certainly as to how we will look at them lies in the scaling of the random variables. In the case of laws of large numbers, the sequences of random variables under consideration have variances that converge to zero, so we get convergence to a constant; typically convergence in probability, which yields weak laws of large numbers but sometimes almost sure convergence, which yields strong laws of large numbers. Central limit theorems are concerned with convergence in distribution. Here the sequence of random variables is scaled to keep the variance of order O(1), so that even in the limit we observe a random variable, rather than a constant. Within each of these broad classes we see that there are many variations in the theme, differing primarily in terms of the assumptions that are made about the existence of moments, their constancy, or in their absence the tail behaviours of the relevant probability distributions.

3.1 Laws of Large Numbers

3.1.1 Weak Laws of Large Numbers

The weak law of large numbers (WLLN) specifies conditions under which the difference $\overline{X}_n - \text{E}\left[\overline{X}_n\right]$ converges in probability to zero, where $\overline{X}_n = n^{-1}\sum_{i=1}^n X_i$ denotes the sample mean. This latter result is often established by demonstrating that $\overline{X}_n - \text{E}\left[\overline{X}_n\right] \xrightarrow{m} 0$. The simplest of these results that we shall consider is that due to Khintchine.

Theorem 3 (Khintchine's Weak Law of Large Numbers)**.** *Let $X_1, X_2, \ldots$ be a sequence of independently and identically distributed (iid) random variables with $\text{E}\left[\overline{X}_n\right] = \mu < \infty$. Then, as $n \to \infty$, $\overline{X}_n \xrightarrow{p} \mu$.* □

The key features of Khintchine's WLLN are that the random variables are iid and that they have a finite mean. We can relax the iidness assumption but to do so comes at a cost. For example, in the next theorem we impose a restriction on the variances of the random variables.

Theorem 4 (Chebyshev's Theorem)**.** *Let $\{X_n, n \geq 1\}$ be a sequence of independent random variables such that $\text{E}[X_i] = \mu_i$ and $\text{Var}[X_i] = \sigma_i^2 < m < \infty$, $i = 1, 2, ..., n$. Then*

for any $\epsilon > 0$

$$\lim_{n\to\infty} \Pr\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \frac{1}{n}\sum_{i=1}^{n}\mu_i\right| < \epsilon\right) = 1. \qquad \square$$

There are lots of variations on this result which we won't explore now. The definitive result for WLLN of this type is due to Kolmogorov.

Theorem 5 (Kolmogorov's Weak Law of Large Numbers)**.** *The sequence of random variables* $\{X_n, n \geq 1\}$ *obeys the WLLN*

$$\operatorname{plim} \overline{X}_n = \mu = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}\left[X_i\right]$$

if and only if

$$\mathrm{E}\left[\frac{\left[\sum_{i=1}^{n} X_i - \sum_{i=1}^{n} \mathrm{E}\left[X_i\right]\right]^2}{n^2 + \left[\sum_{i=1}^{n} X_i - \sum_{i=1}^{n} \mathrm{E}\left[X_i\right]\right]^2}\right] \to 0 \quad \textit{as } n \to \infty. \qquad \square$$

There are very few assumptions here but the requirements of Kolmogorov's condition look fairly fiercesome and, indeed, they are. So although this is a very powerful result, it is extremely difficult to apply in practice because of the difficulties inherent in evaluating the various expectations in the absence of strong distributional assumptions, which is exactly what this sort of result is designed to avoid. However, if you look at the at the various components then it is not as complicated at it looks at first sight. Dividing numerator and denominator through by n^2 reduces the convergence condition to

$$\mathrm{E}\left[\frac{n^{-2}\left[\sum_{i=1}^{n} X_i - \sum_{i=1}^{n} \mathrm{E}\left[X_i\right]\right]^2}{1 + n^{-2}\left[\sum_{i=1}^{n} X_i - \sum_{i=1}^{n} \mathrm{E}\left[X_i\right]\right]^2}\right] \to 0 \quad \text{as } n \to \infty.$$

The expectation of a ratio is not the same as a ratio of expectations but, if we ignore that for a moment and think of taking the expected value of the numerator, it reduces to $\operatorname{Var}\left[\overline{X}\right]$. Similarly, the denominator reduces to $1 + \operatorname{Var}\left[\overline{X}\right]$ under the same loose mathematics. In essence, the condition says that this variance must converge to zero. What causes variances to converge to zero? The tails of the sampling distribution of the sample mean must be becoming thinner as the sample size increases. So, heuristically at least, the condition is telling us just that. Another way of thinking about it is that the variances of the individual sample elements can't explode too severely at any stage because that would cause the variance of the mean to also diverge. Note that the 1 in the denominator simply stops us from dividing by zero at any stage. As for a formal proof of Kolmogorov's WLLN, we will leave that for another time.

In summary, the WLLN says that sample averages converge to population means. It should be recognized that the X_i might be functions of more primitive random variables, e.g. the X_i might be contributions to a log-likelihood function or to a score, which is our context of primary interest in this document. It should also be recognized that we have barely scrapped the surface of this area of study.

3.1.2 Strong Laws of Large Numbers

To provide the flavour of how almost sure convergence provides similar results to Theorem 3 we present here two strong laws of large numbers attributed to Kolmogorov (see Rao, 1973, pp. 114–115):

Theorem 6 (Strong Law of Large Numbers 1)**.** *Let $X_1, X_2, \ldots$ be a sequence of independently and identically distributed (iid) random variables. Then a necessary and sufficient condition that $\overline{X}_n \xrightarrow{a.s.} \mu$ is that $\mathrm{E}[X_1] = \mu < \infty$.*[6] □

Theorem 6 has the same requirements as Theorem 3. So everywhere in the sequel that we claim convergence in probability as a consequence of Khintchine's theorem we might also be claiming almost sure convergence. We won't, however, pursue this issue further.

Theorem 7 (Strong Law of Large Numbers 2)**.** *Let $X_1, X_2, \ldots$ be a sequence of independent random variables with $\mathrm{Var}[X_i] = \sigma_i^2$. If $\sum_{i=1}^{\infty} \sigma_i^2/i^2 < \infty$, then, as $n \to \infty$, $\overline{X}_n - \mathrm{E}[\overline{X}_n] \xrightarrow{a.s.} 0$.* □

This form of the strong law of large numbers again illustrates how stronger results require more stringent conditions. Here we have relaxed the assumption that the terms in the sequence $X_1, X_2, \ldots$ are identically distributed but strong convergence requires a the condition on variances $\sum_{i=1}^{\infty} \sigma_i^2/i^2 < \infty$ which is both more demanding than anything else that we have encountered so far and more difficult to check.[7] This type of condition illustrates why we will restrict attention to convergence in probability in the subsequent discussion.

3.2 Central Limit Theorems

Laws of large numbers typically tell us something about sample averages converging to population moments which are not random. That is, laws of large numbers tell us about sample averages satisfying $\overline{X}_n - \mu = o_p(1)$, which means that differences between sample and population means vanish. Unfortunately such limiting results are not very useful for inference (by which I mean hypothesis testing and the construction of confidence intervals) which involves the distributions of sample statistics. To resolve this problem we adopt *stabilizing transformations* which yield random variables of stochastic order $O_p(1)$; that is, their standard deviation neither converges to zero nor diverges as $n \to \infty$. The simplest such result is the Lindeberg-Levy central limit theorem.

Theorem 8 (Lindeberg-Levy Central Limit Theorem)**.** *Let $X_1, X_2, \ldots$ be iid, with $\mathrm{E}[X_i] = \mu$ and $\mathrm{Var}[X_i] = \sigma^2 < \infty$ for $i = 1, \ldots, n$. Then*

$$Z_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{X_i - \mu}{\sigma} \xrightarrow{d} Z, \tag{2}$$

where $Z \sim \mathrm{N}(0, 1)$. □

Note that it is commonplace to report a result such as (2) as $Z_n \xrightarrow{d} \mathrm{N}(0, 1)$. Although this notation doesn't necessarily make a lot of sense it should be read as meaning that Z_n converges to a random variable with the stated distribution.

[6] Observe that, because the terms in the sequence $X_1, X_2, \ldots$ are *iid* it follows that if $\mathrm{E}[X_1] = \mu < \infty$ then so to do $\mathrm{E}[X_i] = \mu < \infty$ for $i = 2, 3, \ldots$

[7] By way of contrast, a weak law of large numbers would require $n^{-2} \sum_{i=1}^{\infty} \sigma_i^2 < \infty$, which is actually much easier to check than is $\sum_{i=1}^{\infty} \sigma_i^2/i^2 < \infty$.

The transformation from $S(n) = \sum_{i=1}^{n} X_i \to Z_n$ presented in (2) is the stabilizing transformation required to still have a random variable in the limit.[8] In particular, we know that $\mathrm{E}[S(n)] = n\mu$ and that $\mathrm{Var}[S(n)] = n\sigma^2 = O(n)$. We might resolve the diverging standard deviation if we were to divide $S(n)$ by $n^{1/2}$.[9] This yields $\mathrm{Var}\left[n^{-1/2}S(n)\right] = \sigma^2 = O(1)$ and so the variance is controlled. Unfortunately this transformation also yields $\mathrm{E}\left[n^{-1/2}S(n)\right] = n^{1/2}\mu$, which diverges as $n \to \infty$. By first centring the random variable (that is, subtracting away its expected value), we obtain a random variable with zero mean, a mean which will be unchanged by the variance stabilizing transformation.

An alternative form of, or corollary to, the Lindeberg-Levy central limit theorem is in terms of sample means.

Theorem 9 (Sample Mean Form of Lindeberg-Levy Central Limit Theorem)**.** *Let $X_1, X_2, \ldots$ be iid, with $\mathrm{E}[X_i] = \mu$ and $\mathrm{Var}[X_i] = \sigma^2 < \infty$ for $i = 1, \ldots, n$. Then*

$$Z_n = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} = \frac{\sqrt{n}(\overline{X} - \mu)}{\sigma} \xrightarrow{d} \mathrm{N}(0,1). \qquad \square$$

We see that one form of the stabilizing transformation is simply sample mean minus expected value of sample mean, all scaled by the standard deviation of the sample mean.

The Liapunov central limit theorem relaxes the assumption of constant variance,[10] and hence that the elements of the sequence $X_1, X_2, \ldots$ are not identically distributed. As we shall see, this relaxation of assumptions comes at a cost of other restrictions that must be imposed upon the distributions of the random variables.[11]

Theorem 10 (Liapunov Central Limit Theorem)**.** *Let $X_1, X_2, \ldots$ be independent with $\mathrm{E}[X_i] = \mu_i$, $\mathrm{Var}[X_i] = \sigma_i^2$ and $\mathrm{E}[|X_i - \mu_i|^3] = m_i < m < \infty$ for $i = 1, \ldots, n$. Let $C_n = (m_1 + \ldots + m_n)^{1/3}$ and $D_n = (\sigma_1^2 + \ldots + \sigma_n^2)^{1/2}$. If,*

$$\lim_{n\to\infty} C_n/D_n = 0 \qquad \text{(Liapunov Condition)}$$

then

$$Z_n = D_n^{-1} \sum_{i=1}^{n} (X_i - \mu_i) \xrightarrow{d} \mathrm{N}(0,1). \qquad \square$$

The condition on the ratio C_n/D_n is a little more complicated than we had previously but the structure of the central limit theorem is much the same. The sum is a random variable with zero mean and D_n is the standard deviation of the sum and it is the studentized quantity that has the limiting distribution.

An alternative statement of the Liapunov central limit theorem imposes the condition

$$\mathrm{E}\left[|X_n - \mu_n|^{2+\delta}\right], \ \delta > 0$$

[8] These transformations are also known as standardizing or studentizing transformations.

[9] Note that if we had simply scaled by $n^{1/2}$ in (2) then the limiting distribution would have been $\mathrm{N}(0, \sigma^2)$.

[10] Note that Liapunov also appears as Lyapunov, and even Lyapunoff, throughout the literature.

[11] We do not here relax the assumption of independence although that can also be done at the cost of yet more stringent assumptions elsewhere.

rather than the condition on third moments used in our statement. In either case, the role of the condition is to ensure that the tail probabilities of the populations of X are not too thick, so that moments of higher order than two exist.

More powerful than either of the central limit theorems encountered so far is that of Lindeberg and Feller. Here we relax requirements on the existence of moments but at the cost of assumptions about distribution functions.

Theorem 11 (Lindeberg-Feller CLT)**.** *Let $\{X_n\}$ be independent random variables with distribution functions $F_n(x)$, and first two moments $\mathrm{E}[X_n] = \mu_n$ and $\mathrm{Var}[X_n] = \sigma_n^2$. Define*

$$C_n = \left[\sum_{i=1}^{n} \sigma_i^2\right]^{\frac{1}{2}}$$

If

$$\lim_{n\to\infty} \frac{1}{C_n^2} \sum_{i=1}^{n} \int_{|x-\mu_i|>\epsilon C_n} (x-\mu_i)^2 \, \mathrm{d}F_i(x) = 0 \qquad \text{(Lindeberg Condition)}$$

for every $\epsilon > 0$ then

$$Z_n = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} \mathrm{N}(0,1)$$

where $\mu = \mathrm{E}\left[\overline{X}_n\right] = \frac{1}{n}\sum_{i=1}^{n} \mu_i$ and $\sigma^2 = \mathrm{Var}\left[\overline{X}_n\right] = \frac{1}{n^2}\sum_{i=1}^{n} \sigma_i^2$. □

The Lindeberg-Levy requires iidness and existence of variance but no other conditions. The other two CLT's allow for non-identical distributions but require conditions on either higher order moments (Liapunov) or on tail probabilities. If the Liapunov condition is satisfied then so will be the Lindeberg condition, which is as relaxed as this form of central limit theorem can go. Thus, as the Lindeberg condition may be very difficult to verify, the Liapunov result can be thought of as a simpler version to work with for non-iid random variables.

To extend such results to multivariate situations requires a little bit of care.

Theorem 12 (Cramér-Wold Device)**.** *Let $X_1, X_2, \ldots$ be a sequence of k-vectors of random variables. If $c'X_n$ converges in distribution to $c'X$ for every finite $c \in \mathbb{R}^k, c \neq 0$, then $X_n \xrightarrow{d} X$.*

Theorem 13 (Multivariate Central Limit Theorem)**.** *Let $X_1, X_2, \ldots$ be a sequence of k-vectors of random variables. If $c'X_n$ converges to a normally distributed random variable for every non-stochastic k-vector $c \neq 0$ then X_n converges to a multivariate normally distributed vector.*[12] *Specifically, if $c'X_n$ converges in distribution to $\mathrm{N}(0, c'\Omega c)$ for every $c \neq 0$ then the sequence $X_1, X_2, \ldots$ converges in distribution to $\mathrm{N}(0, \Omega)$.*

The central limit theorems that we have seen so far are quite general and can be bent to suit our needs. However, there are forms of central limit theorems that are designed with regression models in mind. We shall finish this section with three such results.

[12] In this case it is not enough to show that each element of X_n converges to a normally distributed random variable, joint normality requires that every random linear combination $c'X_n$ has a normal limiting distribution.

Theorem 14 (Malinvaud (1970, p.251)). *Suppose* $\{\mathbf{u}_n; n = 1, 2, ..., N\}$ *are independent and identically distributed* m*-vector random variables with zero means and finite covariance matrix,* $\mathbf{\Omega}$*. We define*

$$\mathbf{X}_N = N^{-1/2} \sum_{n=1}^{N} \mathbf{A}_n \mathbf{u}_n$$

where $\mathbf{A}_n$ *is an* $p \times m$ *matrix of non-random variables that are uniformly bounded (i.e. the absolute values of the elements of* $\mathbf{A}_n$ *are less than* R *for all* n *and some finite* R*), such that*

$$\lim_{N\to\infty} N^{-1} \sum_{n=1}^{N} \mathbf{A}_n \mathbf{\Omega} \mathbf{A}_n' = \mathbf{V} \text{ is finite.}$$

Then the limiting distribution of $\mathbf{X}_N$ *is* $\mathrm{N}(0, \mathbf{V})$. □

We shall see examples of this result later in these notes. Note that the theorem is able to handle systems of equations. For us it will typically be used in the context of a single linear equation, in which case $\mathbf{u}_n$ reduces to a scalar quantity. Note that if our regressors have different limiting behaviours then we can readily extend the result by replacing the quantity $\mathbf{A}_n$ by $\mathbf{A}_n^* = \mathbf{\Theta}\mathbf{A}_n$ where $\mathbf{\Theta}(N)$ is a (diagonal) matrix with elements dependent on N that admits different rates of scaling.

Unfortunately, Malinvaud's result is unable to deal with the sort of dependence that one finds in many time series models. We address some of these concerns below.

Definition 3. m-Dependence.
Let $x_1, x_2, \ldots$ be a sequence of random variables. This sequence is said to be m-dependent if the two subsequences

$$\{x_1, x_2, \ldots, x_r\} \qquad \{x_s, x_{s+1}, \ldots, x_n\}$$

are independent whenever $r + m < s$.

Remark 1. 0-dependence is equivalent to independence. m-dependence implies that two elements are independent if the difference between their indices is greater than m.

Given this definition, we have the following Liapounov-like result for m-dependent sequences.

Theorem 15 (Hoeffding and Robbins (1948)). *Let* $x_1, x_2, \ldots$ *be a sequence of* m*-dependent random variables with zero mean and with uniformly bounded third absolute moment, so that*

$$\mathrm{E}[x_t] = 0 \qquad \text{and} \qquad \mathrm{E}\left[(\mathrm{abs}(x_t))^3\right] < R < \infty \quad (t = 1, 2, 3 \ldots).$$

Let

$$A_t = \mathrm{E}\left[x_{t+m}^2\right] + 2 \sum_{j=1}^{m-1} \mathrm{E}\left[x_{i+j} x_{i+m}\right]$$

and suppose that

$$\sigma^2 = \lim_{H\to\infty} \frac{1}{H} \sum_{h=1}^{H} A_{t+h}$$

exists and is independent of t*. The random variable*

$$Z_T = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} x_t \xrightarrow{d} \mathrm{N}(0, \sigma^2).$$ □

Finally, an analogue of Malinvaud's central limit theorem that is directed towards regression models.

Theorem 16 (Schönfeld (1971))**.** *Consider the multivariate autoregressive process*

$$\mathbf{y}_t = \boldsymbol{z}_t \mathbf{C}_0 + \sum_{j=1}^{J} \mathbf{y}_{t-j} \mathbf{C}_j + \boldsymbol{\epsilon}_t,$$

where $\mathbf{y}_t$ *is a* $1 \times G$ *vector of the* t*th observation on the* G *dependent variables,* $\mathbf{z_t}$ *is a* $1 \times K$ *vector of the* t*-th observation on the* K *non-stochastic independent variables,* $\boldsymbol{\epsilon}_t$ *is a* $1 \times G$ *disturbance vector,* $\mathbf{C}_0$ *is a* $K \times G$ *coefficient matrix, and* $\mathbf{C}_j$ $(j = 1, \ldots, J)$ *are* $G \times G$ *coefficient matrices. Assume that*

(i) *The* $\boldsymbol{\epsilon}_t'$ *are iid* $\mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma})$*.*

(ii) $\mathrm{abs}\,(Z_{tk}) < M < \infty$*, where* $k = 1, \ldots, K$*;* $t = 1, 2, \ldots$

(iii) *That*

$$\lim_{T \to \infty} \frac{1}{T - \tau} \sum_{t=1}^{T-\tau} \mathbf{z}_t' \mathbf{z}_{t+\tau}$$

exists for all τ*, and*

$$\lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} \mathbf{z}_t' \mathbf{z}_t$$

is non-singular.

(iv) *All roots of the equation*

$$\det \left(\lambda^J \mathbf{I}_G - \sum_{j=1}^{J} \lambda^{J-j} \mathbf{C}_j \right) = 0$$

are less than one in absolute value (or modulus if roots are complex).

Finally, let $\boldsymbol{\epsilon} = [\boldsymbol{\epsilon}_1', \ldots \boldsymbol{\epsilon}_T']'$ *denote the* $T \times G$ *matrix obtained on stacking the* T *disturbance vectors, let*

$$\mathbf{x}_t = [\mathbf{z}_t, \mathbf{y}_{t-1}, \ldots, \mathbf{y}_{t-J}]$$

and let $\mathbf{X}$ *be the* $T \times (K + GJ)$ *matrix of all* T *observations* $\mathbf{x}_t$*. Then*

$$\mathbf{Q} = \operatorname{plim} \frac{\mathbf{X}'\mathbf{X}}{T}$$

is finite and non-singular, and

$$\operatorname{vec} \frac{\mathbf{X}'\boldsymbol{\epsilon}}{\sqrt{T}} \xrightarrow{d} \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma} \otimes \mathbf{Q}).$$ □

Even allowing for dependence of this nature does not deal with all of our problems. In any event, it is not important that you memorize all of these different expression. Rather, the lesson to take from it all is that we can relax assumptions, but it always comes at a cost of additional complexity in establishing the sorts of results that we need to address the problems that arise in the real world.

3.3 The Delta Method

We spent quite a bit of time earlier in the course looking at how to find moments and distributional results for functions of random variables. The *Delta method* is a device for finding approximations to these things when the functions involved are differentiable, as it is based on a Taylor approximation.

1. Suppose, that $n^{1/2}(X_n - \mu) \xrightarrow{d} \mathrm{N}(0, \sigma^2)$, and let g be a continuous differentiable function not involving n, then

$$n^{1/2}[g(X_n) - g(\mu)] \xrightarrow{d} \mathrm{N}(0, [g^{(1)}(\mu)]^2\sigma^2).$$

To understand where this result comes from take a first-order expansion of $g(X_n)$ about $g(\mu)$, thus[13]

$$g(X_n) = g(\mu) + \left[\frac{\partial g(X_n)}{\partial X_n}\right]_{X_n=\mu}(X_n - \mu) + o_p(|X_n - \mu|).$$

Because g does not depend on n, $[\partial g(X_n)/\partial X_n]_{X_n=\mu} = g^{(1)}(\mu) = O(1)$. Rearranging and scaling by $n^{1/2}$ yields

$$n^{1/2}[g(X_n) - g(\mu)] = g^{(1)}(\mu)n^{1/2}(X_n - \mu) + n^{1/2}o_p(|X_n - \mu|).$$

Taking limits of both sides as $n \to \infty$ yields the desired result. Just to be clear on the steps involved let's look at exactly how this works. We will deal with the remainder term first.[14] If g is infinitely differentiable then the remainder is of the form

$$n^{1/2}\left[\frac{g^{(2)}(\mu)}{2!}(X_n - \mu)^2 + \frac{g^{(3)}(\mu)}{3!}(X_n - \mu)^3 + \ldots\right]$$

We see that the typical term of this series is of the form

$$n^{1/2}\frac{g^{(p)}(\mu)}{p!}(X_n - \mu)^p = \underbrace{\frac{g^{(p)}(\mu)}{p!}}_{O(1)}\underbrace{n^{1/2}(X_n - \mu)}_{O_p(1)}\underbrace{(X_n - \mu)^{p-1}}_{o_p(1)=O_p(n^{-(p-1)/2})}, \quad p = 2, 3, \ldots$$

The important point is that given $n^{1/2}(X_n - \mu) = O_p(1)$, by virtue of converging in distribution to something, it follows that $X_n - \mu \xrightarrow{p} 0$. Moreover, powers of $X_n - \mu$

[13] Note that this expression is truly an equality rather than any sort of approximation. At this stage, the term denoted $o_p(|X_n - \mu|)$ is simply

$$o_p(|X_n - \mu|) = g(X_n) - g(\mu) - \left[\frac{\partial g(X_n)}{\partial X_n}\right]_{X_n=\mu}(X_n - \mu).$$

You should also be aware of a notational device that will be used frequently throughout what follows. An expression of the form $[h(x)]_{x=x_0}$ or $h(x)|_{x=x_0}$ should be read as meaning that the function $h(x)$ should be evaluated at the value $x = x_0$. For example, if $h(x) = dg(x)/dx$ then $h(x)|_{x=x_0}$ should be taken to mean that g is first differentiated and then the derivative is evaluated at x_0, rather than first evaluating g at x_0, so that it is no longer a function of x, and then differentiated with respect to x. This allows us to avoid notation like $h(x) = dg(x_0)/dx$ which I always find confusing.

[14] To do this properly require considerably more finesse than will be displayed here where the aim is to provide some intuition.

converge in probability to zero even faster than does $X_n - \mu$, which converges at a rate of $n^{-1/2}$. Hence,

$$\lim_{n\to\infty} n^{1/2} g^{(p)}(\mu)(X_n - \mu)^p / p! = 0$$

because asymptotically it behaves like the product of some real numbers and zero. This leaves[15]

$$n^{1/2}[g(X_n) - g(\mu)] \approx \underbrace{g^{(1)}(\mu)}_{O(1)} \underbrace{n^{1/2}(X_n - \mu)}_{\xrightarrow{d} \mathrm{N}(0,\sigma^2)}$$

and the final result then follows from standard properties of normally distributed random variables when multiplied by constants.

2. Here we repeat the previous example but now we let $X_1, X_2, \ldots$ be a sequence of k-vectors of random variables such that $n^{1/2}(X_n - \mu) \xrightarrow{d} \mathrm{N}(0, \Omega)$, where μ is a k-vector of constants and $\Omega > 0$ is a $k \times k$ matrix of constants. If g is a j-vector of continuous differentiable functions of X_n not involving n, with $j \leq k$. Then

$$n^{1/2}[g(X_n) - g(\mu)] \xrightarrow{d} \mathrm{N}\left(0, G'\Omega G\right), \tag{3}$$

where $G \equiv g^{(1)}(\mu) = [[\partial g(X_n)]'/\partial X_n]_{X_n=\mu}$ is a $k \times j$ matrix of partial derivatives evaluated at $X_n = \mu$.

These two examples illustrate what is known as the Delta method, which is a device for finding the approximate variance of functions of random variables. For our purposes X_n will represent some estimator and the Delta method will be employed to find the variance of functions of the estimator. We shall encounter further examples of this type as we proceed. As a final comment it is worth noting that in practice we will need to estimate the variance of the limiting distribution and so we will typically be estimating $G\Omega G'$ using estimators $\widehat{G}$ and $\widehat{\Omega}$ (wherever they come from) to construct $\widehat{G\Omega G'} = \widehat{G}\widehat{\Omega}\widehat{G}'$. Estimators of this form are commonly referred to as *sandwich estimators* because $\widehat{\Omega}$ is 'sandwiched' between $\widehat{G}$ and $\widehat{G}'$ in constructing the quadratic form. Sandwich estimators are very widely used throughout econometrics.

4 The Asymptotics of the Classical Linear Regression Model and Ordinary Least Squares

4.1 Introduction

We basically have everything in place that we need to construct asymptotic approximations. For what we are going to do the following is sufficient: If some sequence of random variables $X_1, X_2, \ldots$ satisfies

$$\frac{\overline{X}_n - \mu_n}{\sigma_n} \xrightarrow{d} \mathrm{N}(0, 1),$$

[15] At this point we have an approximate result because we have treated the term $n^{1/2} o_p(X_n - \mu)$ as negligible and have ignored it.

where $\mu_n = \mathrm{E}[\overline{X}_n]$ and $\sigma_n^2 = \mathrm{V}[\overline{X}_n]$, then we might approximate the distribution of $\overline{X}_n$ by $\mathrm{N}(\mu_n, \sigma_n^2)$. That is, we obtain an approximating distribution by unravelling the stabilizing transformation require to obtain the limiting distribution. We write

$$\overline{X}_n \underset{a}{\sim} \mathrm{N}(\mu_n, \sigma_n^2).$$

As a first example, suppose that $X_1, X_2, \ldots$ are *iid* with $\mathrm{E}[X_i] = \mu$ and $\mathrm{Var}[X_i] = \sigma^2 < \infty$. Then $\mu_n = \mathrm{E}\left[\overline{X}_n\right] = \mu$, $\sigma_n^2 = \mathrm{Var}\left[\overline{X}_n\right] = \sigma^2/n$, and

$$\frac{\sqrt{n}\left(\overline{X}_n - \mu\right)}{\sigma} \xrightarrow{d} \mathrm{N}(0,1).$$

The asymptotic distribution of $\overline{X}_n$ is

$$\overline{X}_n \underset{a}{\sim} \mathrm{N}\left(\mu, \frac{\sigma^2}{n}\right),$$

which is what we use to approximate the true but unknown sampling distribution of $\overline{X}_n$ in finite samples.

We might expect this approximation to work well if we are sufficiently deep into the sequence of $X_1, X_2, \ldots$ (that is, if n is sufficiently large). This raises the question of when are we sufficiently deep in to the sequence for the asymptotic approximation to work well. There is no answer to this question beyond the sample size being large enough when the approximation works well. Just how large n will have to be depends upon the true distributions of the elements of the sequence $X_1, X_2, \ldots$. In this case, the approximation is actually the exact result if the $X_1, X_2, \ldots$ are a simple random sample from a $\mathrm{N}(\mu, \sigma^2)$ distribution. So one might expect the approximation to work well for relatively small n if the $X_1, X_2, \ldots$ have true distributions which make the sample look like *iid* draws from a Normal population, e.g. if the draws are independent and the true distributions are symmetric about their means. The less like this the data look the greater will the sample size need to be before the approximation works well.

The regression model with X_n's fixed in repeated samples is the most difficult example that we will encounter and the results that are provided are not necessarily the most elegant.[16] Nevertheless, it is presented as a device to demonstrate a number of different ideas. First amongst these is the way increasingly restrictive assumptions cascade through the problem as we seek to establish consistency and then asymptotic normality for the least squares estimators of the regression coefficients and then the unbiased estimator of the variance, respectively. The more we wish to know, the more information the problem demands. Second, we take the opportunity to demonstrate the use of the Liapunov central limit theorem. Third, we demonstrate the use of the Cramér-Wold device for establishing the limiting distribution of a sequence of random vectors.[17] Hereafter we shall simply

[16] The treatment in Amemiya (1985, Section 3.5) is superior but is also much more demanding in its presumed knowledge, especially of matrix algebra. The cost of not using these more sophisticated arguments is that the derivations presented here are much more of a grind that they might otherwise be. Such is life.

[17] Harald Cramér (1893–1985), a great Swedish mathematician, actuary, and statistician. As well as his own substantial contributions to probability theory, his book Cramér (1946) is an absolute classic, he supervised a number of students who went on to make substantial contributions to probability theory including Herman Wold (1908–1992), Kai Lai Chung (1917–2009), and Ulf Grenander (1923 – 2016). Uncommonly he also became a substantial figure in University administration in Sweden in the latter

assert that such arguments can be applied and then appeal to multivariate central limit theorems as required. Fourth, we have the opportunity to practice arguments that require Big O and Little o notation. Finally, in analyzing the large-sample properties of s^2 we get to explore a sample statistic which is a more complicated function of data than a weighted sum, which is what the least squares estimators are. As a by-product we also establish a result that we use later in the analysis of maximum likelihood. As you can see there is a lot in this example and it is easy to be overwhelmed by the details. The details are provided only for those who are interested. If that is not you then please focus on the broader issues described above.

4.2 The Assumptions of the Model

To begin, we shall characterize the model according to the following assumptions:

A.1 The true model is

$$y_i = x_i'\beta + u_i, \qquad i = 1, \ldots, n$$

or, in matrix notation,

$$\tilde{y}_n = X_n\beta + \tilde{u}_n$$

where $\tilde{y}_n = [y_1, \ldots, y_n]'$ is a random n-vector, $X_n = [x_i, \ldots, x_n]'$ is an $n \times k$ matrix with full column rank,[18] and $\tilde{u}_n = [u_i, \ldots, u_n]'$ is also a random n-vector. β is a k-vector of parameters that need to be estimated. □

A.2 We will assume that the sequences of explanatory variables neither diverge nor converge to zero which, perhaps surprisingly, is more of a problem. We write this assumption as

$$\operatorname*{plim}_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} x_i x_i' = \operatorname*{plim}_{n\to\infty} \frac{1}{n} X_n'X_n = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}\left[x_i x_i'\right] = Q,$$

where Q is a finite positive definite matrix. □

A.3 Given X_n, the elements of $\tilde{u}_n$ have (i) zero mean, so that $\mathrm{E}\left[\tilde{u}_n \mid X_n\right] = 0$ and hence $\mathrm{E}\left[X_n'\tilde{u}_n\right] = 0$, (ii) constant variance, so that $\mathrm{Var}\left[u_i \mid X_n\right] = \sigma^2$, $\sigma^2 < \infty$, and (iii) are independent of one another. Together, these elements imply that $\mathrm{Var}\left[\tilde{u}_n \mid X_n\right] = \mathrm{E}\left[\tilde{u}_n\tilde{u}_n' \mid X_n\right] = \sigma^2 I_n$. □

Assumption A.3 is very strong and, indeed, stronger than we need. For a start, if there is any shared sources of randomness underlying both the disturbance term and the explanatory regressors then exogeneity of the regressors fails and we are unable to condition on the X's in the way that we do. Even given exogeneity of the X_n, the assumptions of conditional homoskedasticty and, especially, the conditional independence of the disturbances are very strong. In any event, our aim is to illustrate the ideas rather than to

part of his career. Wold was a Swedish econometrician who was born in Norway. In addition to his own substantial contributions to the literature he also supervised the PhD of the great Peter Whittle, among others. Chung wrote many books on probability theory, e.g., Chung (1968) and its subsequent editions, and made substantial contributions to the theory of Brownian motion and Markov chains. Grenander was another Swedish mathematician who made substantial contributions to probability theory, notably in the areas of probability theory, stochastic processes and time series.

[18] It is important to recognize that means that $n \geq k$ in everything that follows.

provide the most general treatment possible. As we saw in Section 3.2, it is possible to relax the various assumptions but at the cost of increasingly complicated limiting results. So we will live with Assumption A.3 for now but will be conscious of the fact that there is more that can be done.

The OLS estimator for β, denoted $\hat{\beta}_n$, is

$$\hat{\beta}_n = \left[\sum_{i=1}^{n} x_i x_i'\right]^{-1} \sum_{i=1}^{n} x_i y_i = \beta + \left[\sum_{i=1}^{n} x_i x_i'\right]^{-1} \sum_{i=1}^{n} x_i u_i$$

or

$$\hat{\beta}_n = (X_n' X_n)^{-1} X_n' \tilde{y}_n = \beta + (X_n' X_n)^{-1} X_n' \tilde{u}_n \tag{4}$$

The final members of these expressions are the most important representations for what follows. Furthermore, we have the usual unbiased estimator for σ^2

$$s_n^2 = \frac{1}{n-k} \sum_{i=1}^{n} (y_i - x_i' \hat{\beta}_n)^2 = \frac{(\tilde{y}_n - X_n \hat{\beta}_n)'(\tilde{y}_n - X_n \hat{\beta}_n)}{n-k}.$$

We shall proceed using matrix notation because it is a little cleaner than the sigma notation, however, all the relevant expressions have been provided in sigma notations and so you can make the relevant substitutions for yourself if desired.

4.3 On the Consistency of the OLS Estimator

Now, using our various rules for probability limits (see Theorem 1 on page 10)

$$\begin{aligned}
\operatorname*{plim}_{n\to\infty} \hat{\beta}_n &= \operatorname{plim} \beta + \operatorname{plim} (X_n' X_n)^{-1} X_n' \tilde{u}_n \\
&= \beta + \operatorname{plim} \left[\left(\frac{X_n' X_n}{n} \right)^{-1} \left(\frac{X_n' \tilde{u}_n}{n} \right) \right] \\
&= \beta + \operatorname{plim} \left(\frac{X_n' X_n}{n} \right)^{-1} \operatorname{plim} \left(\frac{X_n' \tilde{u}_n}{n} \right) \\
&= \beta + Q^{-1} \operatorname{plim} \left(\frac{X_n' \tilde{u}_n}{n} \right)
\end{aligned}$$

The easiest way to evaluate the remaining probability limit is by establishing mean square convergence. Specifically we see that, for all $n \geq k$,

$$\mathrm{E}\left[n^{-1} X_n' \tilde{u}_n\right] = n^{-1} \mathrm{E}_{X_n}\left[X_n' \mathrm{E}\left[\tilde{u}_n \mid X_n\right]\right] = n^{-1} \mathrm{E}_{X_n}\left[X_n' 0\right] = n^{-1} \mathrm{E}_{X_n}\left[0\right] = 0,$$

where the first equality is an application of the law of iterated expectations, and

$$\begin{aligned}
\operatorname{Var}\left[n^{-1} X_n' \tilde{u}_n\right] &= \mathrm{E}\left[n^{-2} X_n' \tilde{u}_n \tilde{u}_n' X_n\right] = n^{-2} \mathrm{E}_{X_n}\left[X_n' \mathrm{E}\left[\tilde{u}_n \tilde{u}_n' \mid X_n\right] X_n\right] \\
&= n^{-2} \sigma^2 \mathrm{E}_{X_n}\left[X_n' X_n\right].
\end{aligned}$$

Combining these results, we see that

$$\lim_{n\to\infty} \mathrm{E}\left[n^{-1} X_n' \tilde{u}_n\right] = \lim_{n\to\infty} 0 = 0$$

and that

$$\lim_{n\to\infty} \operatorname{Var}\left[n^{-1}X_n'\tilde{u}_n\right] = \lim_{n\to\infty} n^{-2}\sigma^2 \operatorname{E}_{X_n}\left[X_n'X_n\right] = \lim_{n\to\infty}\left(\frac{\sigma^2}{n}\right)\lim_{n\to\infty}\left(\frac{\operatorname{E}_{X_n}\left[X_n'X_n\right]}{n}\right)$$
$$= 0 \times Q = 0.$$

Thus, we have established that $n^{-1}X_n'\tilde{u}_n \xrightarrow{m} 0$, which implies that

$$\operatorname{plim} n^{-1}X_n'\tilde{u}_n = 0$$

and hence that $\operatorname{plim}\hat{\beta}_n = \beta$. That is, $\hat{\beta}_n$ is consistent for β.

4.4 The Asymptotic Distribution of the OLS Estimator

Given that $\hat{\beta}_n$ is consistent for β it follows that it has a degenerate distribution in the limit, i.e. a distribution that collapses to a spike at β. Therefore, in order to obtain an asymptotic distribution for $\hat{\beta}_n$ we need to determine a suitable stabilizing transformation which will yield a non-degenerate limiting distribution. Such transformations typically involve centring the random variable of interest, to obtain a quantity with zero mean, and then scaling this difference appropriately. Under our assumptions, $\operatorname{E}\left[\hat{\beta}_n\right] = \beta$ and so the quantity to scale is $\hat{\beta}_n - \beta$.[19] Recall that the order in probability of a random variable is its standard deviation and so we must find this quantity. Now,

$$\begin{aligned}
\operatorname{Var}\left[\hat{\beta}_n\right] &= \operatorname{E}\left[(\hat{\beta}_n - \beta)(\hat{\beta}_n - \beta)'\right] \\
&= \operatorname{E}_{X_n}\left[(X_n'X_n)^{-1}X_n'\operatorname{E}\left[\tilde{u}_n\tilde{u}_n' \mid X_n\right]X_n(X_n'X_n)^{-1}\right] \\
&= \operatorname{E}_{X_n}\left[(X_n'X_n)^{-1}X_n'(\sigma^2\mathbf{I}_n)X_n(X_n'X_n)^{-1}\right] \\
&= \sigma^2\operatorname{E}_{X_n}\left[(X_n'X_n)^{-1}\right],
\end{aligned}$$

for all $n \geq k$. But, by assumption A.2, the term

$$\operatorname{E}_{X_n}\left[n^{-1}X_n'X_n\right] = O(1) \implies \operatorname{E}_{X_n}\left[X_n'X_n\right] = O(n) \implies \operatorname{E}_{X_n}\left[(X_n'X_n)^{-1}\right] = O(n^{-1}),$$

which suggests that a suitable scale factor is $n^{1/2}$, giving

$$\operatorname{Var}\left[n^{1/2}\hat{\beta}_n\right] = \sigma^2(n^{-1}X_n'X_n)^{-1} = O(1).$$

That is, from (4), the standardized quantity of interest is

$$n^{1/2}\left(\hat{\beta}_n - \beta\right) = n^{1/2}\left(X_n'X_n\right)^{-1}X_n'\tilde{u}_n = \left(n^{-1}X_n'X_n\right)^{-1}n^{-1/2}X_n'\tilde{u}_n. \tag{5}$$

In the notation of Theorem 14 we have the following substitutions: $n \equiv i$, $N \equiv n$, $\mathbf{A}_n \equiv x_i$ where x_i' is the ith row of X_n, $\mathbf{u}_n \equiv u_i$ is the ith element of $\tilde{u}_n$ and $\mathbf{\Omega} \equiv \sigma^2$, so that

$$\mathbf{X}_N \equiv n^{-1/2}\sum_{i=1}^{n} x_i u_i$$

[19] It may be the case that you don't know the expected value of your random variable but do know its probability limit. In such cases you would work with the difference between the random variable and its probability limit.

is a k-vector. Then

$$\lim_{n\to\infty N} N^{-1}\sum_{n=1}^{N}\mathbf{A}_n\mathbf{\Omega}\mathbf{A}_n' \equiv \lim_{n\to\infty} n^{-1}\sum_{i=1}^{n} x_i(\sigma^2)x_i' = \sigma^2 \lim_{n\to\infty} n^{-1}\sum_{i=1}^{n} x_i x_i'$$
$$= \sigma^2 \lim_{n\to\infty} n^{-1}X_n'X_n = \sigma^2 Q,$$

so that

$$n^{-1/2}\sum_{i=1}^{n} x_i u_i \xrightarrow{d} W \sim \mathrm{N}\left(0, \sigma^2 Q\right).$$

Therefore, from (5), as $n \to \infty$

$$n^{1/2}\left(\hat{\beta}_n - \beta\right) \to Q^{-1}W \sim \mathrm{N}\left(0, \sigma^2 Q^{-1}Q(Q^{-1})'\right) = \mathrm{N}\left(0, \sigma^2 Q^{-1}\right),$$

where the final equality follows because Q, and hence, Q^{-1} is symmetric. Given this well-defined limiting result, for infinitely large sample, we can approximate the sampling distribution of $\hat{\beta}_n$ in finite samples by unwinding the stabilizing transformation. That is, by dividing both side by $n^{1/2}$ and then adding β to both sides, to obtain

$$\hat{\beta}_n \underset{a}{\sim} \mathrm{N}\left(\beta, \sigma^2 n^{-1}Q^{-1}\right).$$

Alternatively, in accord with Theorem 12, let us consider

$$n^{-1/2}c'X_n'\tilde{u}_n = n^{-1/2}\sum_{i=1}^{n} v_i, \tag{6}$$

where $c \neq 0$ is an arbitrary, non-stochastic k-vector, so that $v_i = c'x_i u_i$ is a scalar quantity. Now, from the conditional independence of the u_i given X_n it follows that the v_i must also be conditionally independent given X_n. We see that $\mathrm{E}[v_i] = 0$ but that $\mathrm{Var}[v_i] = \sigma^2 c'x_i x_i' c$. Hence the v_i are not identically distributed. Nevertheless, we might appeal to the Liapunov central limit theorem if we are prepared to extend our Assumption A.3 to

A.3′ The u_i, $i = 1, \ldots, n$ are conditionally independent random variables, given the X_n, with $\mathrm{E}[u_i \mid X_n] = 0$, $\mathrm{Var}[u_i \mid X_n] = \sigma^2$, and third absolute moment $0 \leq \mathrm{E}[|u_i|^3] = m_i \leq m < \infty$ for $i = 1, \ldots, n$.

Under this revised assumption, which is stronger than our original, we now have independence of the disturbances rather than just a lack of correlation.[20] Observe that $\mathrm{E}[|v_i|^3] \leq |c'x_i|^3 m$. By assumption we know that $m = O(1)$ and so we need to examine the behaviour of $|c'x_i|^3$. Consider first $|c'x_i|$. By the Cauchy-Schwarz inequality we have[21]

$$|c'x_i| \leq \sqrt{c'c}\,\sqrt{x_i'x_i}.$$

Now, $c'c = c_1^2 + \cdots + c_k^2$, where c_j denotes the jth element of c. As c is simply an arbitrary non-zero vector of finite constants and it follows that each c_j^2 is also some finite constant

[20] A lack of correlation is only the same as independence if the population is Normal.
[21] The Cauchy-Schwarz inequality is discussed in Appendix B.

and so, being a sum of k finite constants, it follows that $c'c = O(1)$. That is, $c'c$ is just some finite constant.

Next consider the term $x_i'x_i = x_{i1}^2 + \cdots + x_{ik}^2$, where x_{ij} denotes the jth element of the ith row of X_n. By assumption $\text{plim}(n^{-1}X_n'X_n - Q) = 0$, with $Q = O(1)$. Recall that $X_n'X_n = \sum_{i=1}^n x_i x_i'$, where

$$x_i x_i' = \begin{bmatrix} x_{i1}^2 & x_{i1}x_{i2} & \dots & x_{i1}x_{ik} \\ x_{i2}x_{i1} & x_{i2}^2 & \dots & x_{i2}x_{ik} \\ \vdots & & \ddots & \vdots \\ x_{ik}x_{i1} & \dots\dots\dots & & x_{ik}^2 \end{bmatrix}.$$

Furthermore, from Assumption A.2, a sum of such matrices is of order $O(n)$ — so that their average converges to a finite positive definite matrix — which suggests that the elements of the matrix $x_i x_i' = O(1)$. Then $x_i'x_i = O(1)$, being the sum of the k terms on the leading diagonal of the matrix. Combining these results we see that

$$|c'x_i| \leq \sqrt{c'c}\sqrt{x_i'x_i} = O(1^{1/2}) \times O(1^{1/2}) = O(1)$$

and, consequently, we have $|c'x_i|^3 = O(1^3) = O(1)$. By exactly the same arguments we may conclude that $\text{Var}[v_i] = O(1)$.

The terms in our sum (6) are of the form $n^{-1/2}v_i$. Before we apply the Liapunov CLT (Theorem 10) we should check the Liapunov condition. Recalling that $\text{E}\left[n^{-1/2}v_i\right] = 0$, define

$$C_n = \sum_{i=1}^n \text{E}\left[\left|n^{-1/2}v_i\right|^3\right] \leq n^{-3/2}m\sum_{i=1}^n |c'x_i|^3 = O\left(n^{-1/2}\right),$$

where the final equality follows on noting that $n^{-3/2} = O(n^{-3/2})$, $m_3 = O(1)$, and $\sum_{i=1}^n |c'x_i|^3 = O(n)$, being the sum of n terms, each of order $O(1)$. Similarly, define[22]

$$D_n = \sum_{i=1}^n \text{E}\left[\left|n^{-1/2}v_i\right|^2\right] = n^{-1}\sigma^2\sum_{i=1}^n |c'x_i|^2 = O(1).$$

It follows that

$$\lim_{n\to\infty}\frac{C_n^{1/3}}{D_n^{1/2}} = \lim_{n\to\infty}\frac{O\left([n^{-1/2}]^{1/3}\right)}{O([1]^{1/2})} = \lim_{n\to\infty}\frac{O\left(n^{-1/6}\right)}{O(1)} = 0.$$

Given that this condition is satisfied, Theorem 10 implies that

$$n^{-1/2}D_n^{-1}\sum_{i=1}^n v_i \xrightarrow{d} \text{N}(0,1),$$

or equivalently, on noting that $\lim_{n\to\infty} D_n^2 = \sigma^2 c'Qc$

$$n^{-1/2}\sum_{i=1}^n v_i \xrightarrow{d} \text{N}(0, \sigma^2 c'Qc).$$

[22] Observe that $n^{-1}\sum_{i=1}^n |c'x_i|^2 = c'(n^{-1}X_n'X_n)c$, In general, if $A = O(1)$ is a $k \times k$ matrix then

$$c'Ac = \sum_{i=1}^k\sum_{j=1}^k c_i c_j a_{i,j} = O(k^2).$$

On noting that our result holds for all non-stochastic k-vectors $c \neq 0$, we see that we have met the requirements of Theorem 13 and have, consequently, established that in the model described by assumptions A.1, A.2, and A.3′, that

$$n^{-1/2} X_n' \tilde{u}_n \xrightarrow{d} \mathrm{N}(0, \sigma^2 Q).$$

With this result in hand we can now return to (5) and solve our initial problem. Thus,

$$n^{1/2}\left(\hat{\beta}_n - \beta\right) = \underbrace{\left(n^{-1} X_n' X_n\right)^{-1}}_{\xrightarrow{p} Q^{-1}} \underbrace{n^{-1/2} X_n' \tilde{u}_n}_{\xrightarrow{d} \mathrm{N}(0,\sigma^2 Q)} \xrightarrow{d} \mathrm{N}\left(0, \sigma^2 Q^{-1}\right). \tag{7}$$

As before, this yields the asymptotic approximation

$$\hat{\beta}_n \underset{a}{\sim} \mathrm{N}\left(\beta, \sigma^2 n^{-1} Q^{-1}\right),$$

albeit with quite a bit more work.[23] In practice, we typically estimate σ^2 by s_n^2 and $n^{-1}Q^{-1} = n^{-1}(\lim_{n\to\infty} n^{-1} X_n' X_n)^{-1}$ by $(X_n' X_n)^{-1}$. That is,

$$\hat{\beta}_n \underset{a}{\sim} \mathrm{N}\left(\beta, \sigma^2 (X_n' X_n)^{-1}\right),$$

which is exactly what we would have got if we had assumed that the true disturbance population was Normal, which is why you learn the model with that assumption in the first place. The difference here is that we understand it to only be approximately true, where the approximation is predicated on some very strong assumption. In order to make the approximation operational we will probably need to replace σ^2 by some estimate.[24] Typically, people will use the unbiased estimator s_n^2 for this purpose, making the final approximation

$$\hat{\beta}_n \underset{a}{\sim} \mathrm{N}\left(\beta, s_n^2 (X_n' X_n)^{-1}\right).$$

Now, s_n^2 is unbiased in finite samples but our approximation is based on limiting arguments and so it makes sense to learn how s_n^2 behaves as the sample size becomes infinitely large, to see if this is a really good idea. It is to this that we turn our attention next.

4.5 On the Consistency of the Unbiased Estimator of the Disturbance Variance

To begin, let us write

$$s_n^2 = \frac{1}{n-k}\sum_{i=1}^{n}(y_i - x_i'\hat{\beta}_n)^2 = \frac{\tilde{u}_n'[I_n - X_n(X_n'X_n)^{-1}X_n']\tilde{u}_n}{n-k},$$

where the second equality is obtained on substituting $(X_n'X_n)^{-1}X_n'\tilde{y}_n$ for $\hat{\beta}_n$. Observe that we can write

$$\begin{aligned} s_n^2 &= \frac{1}{n-k}\left[\tilde{u}_n'\tilde{u}_n - \tilde{u}_n' X_n (X_n'X_n)^{-1} X_n'\tilde{u}_n\right] \\ &= \frac{1}{n-k}\left[\tilde{u}_n'\tilde{u}_n - \frac{\tilde{u}_n' X_n}{n^{1/2}}\left(\frac{X_n'X_n}{n}\right)^{-1}\frac{X_n'\tilde{u}_n}{n^{1/2}}\right]. \end{aligned}$$

[23] For this reason one would typically adopt the analysis based on Theorem 14, but it never hurts to see a lower level analysis at least once.

[24] I would argue that if you know σ^2 then you probably know enough else that you don't need to worry about approximations in the first place.

We have seen that $n^{-1/2}X_n'\tilde{u}_n \xrightarrow{d} \mathrm{N}(0, \sigma^2 Q)$ and that $\mathrm{plim}_{n\to\infty} n^{-1}X_n'X_n = Q$. Hence

$$\frac{\tilde{u}_n'X_n}{n^{1/2}}\left(\frac{X_n'X_n}{n}\right)^{-1}\frac{X_n'\tilde{u}_n}{n^{1/2}} \xrightarrow{d} Z'Q^{-1}Z,$$

where $Z \sim \mathrm{N}(0, \sigma^2 Q)$. From the properties of normally distributed random variables we know that if the k-vector $Z \sim \mathrm{N}(0, \Omega)$ then $Z'\Omega^{-1}Z \sim \chi^2_k$.[25] Furthermore, in Section 2.5 we saw that if $X_n \xrightarrow{d} X$ then $g(X_n) \xrightarrow{d} g(X)$. Together these results tell us that

$$W_n = \frac{\tilde{u}_n'X_n}{n^{1/2}}\left(\sigma^2\frac{X_n'X_n}{n}\right)^{-1}\frac{X_n'\tilde{u}_n}{n^{1/2}} \xrightarrow{d} \chi^2_k = O_p(1).$$

Now, the term that we actually need to analyze is $\sigma^2 W_n/[(n-k)] = O_p(n^{-1})$, which clearly vanishes as $n \to \infty$. This leaves

$$\frac{\tilde{u}_n'\tilde{u}_n}{n-k} = \frac{1}{n-k}\sum_{i=1}^n u_i^2 = \frac{n}{n-k} \times \frac{1}{n}\sum_{i=1}^n u_i^2.$$

Recognizing that $n/(n-k) \to 1$ as $n \to \infty$ we are left with a simply average of squared disturbance terms. By assumption, $\mathrm{E}[u_i^2] = \sigma^2$ and so

$$\mathrm{E}\left[\frac{1}{n}\sum_{i=1}^n u_i^2\right] = \frac{1}{n}\sum_{i=1}^n \sigma^2 = \sigma^2 < \infty.$$

By Khintchine's Theorem (Theorem 3), this is sufficient to establish that

$$\frac{1}{n}\sum_{i=1}^n u_i^2 \xrightarrow{p} \sigma^2,$$

which demonstrates that $s_n^2 \xrightarrow{p} \sigma^2$; that is, s_n^2 is consistent for σ^2.

4.6 The Asymptotic Distribution of the Unbiased Estimator of the Disturbance Variance

One important feature of our demonstration of the consistency of s_n^2 for σ^2 is that it demonstrated the fact that the term

$$(n-k)^{-1}\tilde{u}_n'X_n(X_n'X_n)^{-1}X_n'\tilde{u}_n = O_p(n^{-1})$$

is of smaller order in probability than is $(n-k)^{-1}\tilde{u}_n'\tilde{u}_n$ and so is asymptotically irrelevant.[26] From result 4 on page 14 we see that the limiting distribution of σ^2 will correspond to that of $(n-k)^{-1}\tilde{u}_n'\tilde{u}_n$. The obvious thing to do here is to appeal to the Lindeberg-Levy central limit theorem which tells us that

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n \frac{u_i^2 - \sigma^2}{\sqrt{\mathrm{Var}[u_i^2]}} = \frac{1}{\sqrt{n}}\sum_{i=1}^n \frac{u_i^2 - \sigma^2}{\sqrt{\mathrm{E}[u_i^4] - (\sigma^2)^2}} \xrightarrow{d} \mathrm{N}(0,1)$$

[25] See Appendix C of the Normality handout for this and a number of other similar results.

[26] So why is this term there in the first place? Obviously it matters in finite (or small) samples. It is essentially there as a bias correction.

The only problem with such an appeal is that we don't know anything about $\mathrm{E}[u_i^4]$. That is, in order to make statements about variances we need to make assumptions about fourth moments of the disturbances.[27] Similarly, in order to control sums of cubed random variables we would need to control their sixth moments, and so on. For what we are doing here it is necessary to modify our third assumption along the following lines.

A.3′ The u_i, $i = 1, \ldots, n$ are independent random variables with $\mathrm{E}[u_i] = 0$, $\mathrm{Var}[u_i] = \sigma^2$, and fourth moment $0 < \mathrm{E}[u_i^4] = m_{4i} \leq m_4 < \infty$.

With this extended condition we see that

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{u_i^2 - \sigma^2}{\sqrt{m_4 - (\sigma^2)^2}} = \frac{\sqrt{n}(\tilde{s}_n^2 - \sigma^2)}{\sqrt{m_4 - (\sigma^2)^2}} \xrightarrow{d} \mathrm{N}(0, 1),$$

where

$$\tilde{s}_n^2 = \frac{1}{n} \sum_{i=1}^{n} u_i^2 = \frac{n-k}{n} \frac{1}{n-k} \sum_{i=1}^{n} u_i^2 = \frac{n-k}{n} s_n^2.$$

As $(n-k)/n \to 1$ it follows from $\tilde{s}^2$ will have the same limiting distribution as does

$$s_n^2 = (n-k)^{-1} \sum_{i=1}^{n} u_i^2.$$

That is,

$$s_n^2 \underset{a}{\sim} \mathrm{N}\left(\sigma^2, \frac{m_4 - (\sigma^2)^2}{n}\right). \tag{8}$$

As an aside, and for future reference, if we were to assume that the disturbances were normal then $m_4 = 3\sigma^4$ and (8) becomes

$$s_n^2 \underset{a}{\sim} \mathrm{N}\left(\sigma^2, \frac{2\sigma^4}{n}\right). \tag{9}$$

4.7 One More Observation

Assumption A.2 is that $\lim_{n\to\infty} n^{-1} \mathrm{E}[X_n' X_n] = Q = O(1)$, with $Q > 0$. This assumption implies that the sequences of observations in each of the columns of X_n (explanatory variables, if you prefer) are $O(1)$ for all $i = 1, 2, \ldots$ This precludes, amongst other things, the presence of time trends. We can extend our analysis to account for certain differences in behaviour as follows.

A.2′ Let E_n be a diagonal matrix with typical element n^{p_i}, $i = 1, \ldots, k$, on its leading diagonal. That is, $E_n = \mathrm{diag}(n^{p_1}, \ldots, n^{p_k})$. Then assume that

$$\operatorname*{plim}_{n\to\infty} E_n^{-1/2} X_n' X_n E_n^{-1/2} = Q,$$

where Q is a finite positive definite matrix.

Then, for example, the asymptotic approximation to the distribution of $\hat{\beta}_n$ is obtained from the limiting distribution of $E_n^{-1/2}(\hat{\beta}_n - \beta)$. We see that setting $p_1 = \ldots = p_k = 1/2$, so that $E_n = n^{1/2} I_n$ yields our earlier results. This extension certainly doesn't cover every possible circumstance but does illustrate one way in which these results might be extended.

[27] In essence we are making assumptions about the variances of the squared terms, in exactly the same way as we earlier made assumptions about population variances in order to construct asymptotic arguments about sample means.

5 Time Series Models

5.1 Notation

Time series models typically complicate asymptotic analysis because they are models which seek to explain dependence across time and so violate the independence assumption that has underpinned what has gone before. A serious treatment of this topic is well beyond the scope of what we are attempting here, indeed it could easily form the contents of a very large book (and more). Our aim in this section is simply to give a flavour of some relatively simple results from this literature to illustrate some of the problems which arise. In what follows we shall on occasion make mention of autoregressive-moving average ARMA(p,q) processes of the form

$$\underbrace{Y_t - \phi_1 Y_{t-1} - \cdots - \phi_p Y_{t-p}}_{\text{autoregressive component}} = \underbrace{\epsilon_t + \theta_1 \epsilon_{t-1} + \theta_q \epsilon_{t-p}}_{\text{moving average component}}, \tag{10}$$

or

$$\phi(L) Y_t = \theta(L) \epsilon_t \tag{11}$$

where

$$\begin{aligned} \phi(L) &= 1 - \phi_1 L - \cdots - \phi_p L^p, \\ \theta(L) &= 1 + \theta_1 L + \theta_q L^q, \end{aligned}$$

with L denoting the lag operator which is defined by the transformation $LY_t = Y_{t-1}$. Similarly, $LY^{t-1} = Y_{t-2}$, but $LY_{t-1} = L(LY_t) = L^2 Y_t = Y_{t-2}$, and so on. If we formally define $L^0 = 1$, so that $L^0 Y_t = Y_t$ then, in general,

$$L^r Y_t = Y_{t-r}, \quad r = 0, 1, 2, 3, \ldots \tag{12}$$

In what follows we will treat L like any regular scalar quantity and subject it to algebraic manipulations in the same way that we would any other real quantity. Note that L is closely related to the first difference operator Δ, which is defined according to

$$\Delta Y_t = Y_t - Y_{t-1} = (1 - L) Y_t. \tag{13}$$

In much of what follows we shall restrict attention to the special case of a first-order autoregressive model, i.e., an AR(1) model, as this allows us to illustrate the relevant ideas without the complications that arise from dealing with more general processes. That said, it should be taken as given that there is much, much more to learn about time series models than is discussed here.

5.2 Stationary Autoregressive Models

Consider the model

$$Y_t = \phi Y_{t-1} + \epsilon_t, \tag{14}$$

where Y_t denotes the variable that we are attempting to model, α is some finite constant, and we shall assume that:[28]

[28] The normality is stronger than we need, which is really just the independence across time and the constancy of moments, with zero mean.

1. ϵ_t and ϵ_{t-s} are jointly normally distributed for all t and s,
2. $\epsilon_t \sim \mathrm{N}(0, \sigma^2)$ for all t, and
3. $\mathrm{Cov}\left[\epsilon_t, \epsilon_{t-s}\right] = 0$ for all $s \neq 0$.

Let us look at each of these assumptions in turn. The assumption of joint normality implies that the marginal distributions of the ϵ_t are Normal for all values of t. The second assumption provides an explicit statement of these marginal distributions. The most important aspect of this assumption is that the distribution does not depend upon t, so that all of the ϵ_t have the same marginal distribution. Finally we have assumed that any pair $\{\epsilon_t, \epsilon_{t-s}\}$, $s \neq 0$, are uncorrelated. When coupled with the assumption of joint normality, the lack of correlation amounts to an assumption of independence. Note that a lack of correlation does not imply independence with any other joint distribution.

The assumptions that we have made about the ϵ_t are those of a Gaussian white noise process. Such processes satisfy the assumptions underlying our earlier asymptotic results but we are not interested in models of the ϵ_t, rather we are interested in models of the Y_t. In this Section we restrict attention to *stationary* models, so it makes sense to provide a definition of stationarity. There are two definitions in common use. The first such definition is that of *strict stationarity*. Under this definition, it must be the case that the joint probability of a set of r observations at times $t_1, t_2, \ldots, t_r$ is identical to the joint probability of a set of r observations at times $t_1 + \tau, t_2 + \tau, \ldots, t_r + \tau$, for arbitrary integer τ. This is a very stringent definition that is difficult to check without strong distributional assumptions. Much more common will be the definition adopted here of *weak stationarity* or *covariance stationarity*, which is defined as follows for all values of t:

$$\mathrm{E}\left[Y_t\right] = \mu, \tag{15a}$$
$$\mathrm{Var}\left[Y_t\right] = \gamma(0), \tag{15b}$$
$$\mathrm{Cov}\left[Y_t, Y_{t-s}\right] = \gamma(s), \quad s = 1, 2, 3, \ldots, \tag{15c}$$

where the $\gamma(s)$ are the autocovariance functions of Y_t with itself and its own past.[29] The key features of these functions in this context is that they are not functions of the index t, and so they are the same at any point in time, and that they are all finite. This last requirement has implications for the coefficients $\phi_1, \ldots, \phi_m$. Observe that we can back substitute to obtain

$$\begin{aligned} Y_t &= \phi(\phi Y_{t-2} + \epsilon_{t-1}) + \epsilon_t \\ &= \phi^2(\phi Y_{t-3} + \epsilon_{t-2}) + \epsilon_t + \phi\epsilon_{t-1} \\ &\vdots \\ &= \phi^{r+1} Y_{t-r-1} + \sum_{j=0}^{r} \phi^j \epsilon_{t-j}. \end{aligned} \tag{16}$$

[29] The autocovariance functions can be standardized by the variance to produce an autocorrelation function

$$\rho(s) = \gamma(s)/\gamma(0), \quad s = 0, \pm 1, \pm 2, \ldots$$

We note that $\rho(0) = 1$ and that the autocovariances are symmetric about $s = 0$, so that $\gamma(s) = \gamma(-s)$. The *correlogram* plots the sample analogues of the autocorrelation function, for positive values of s. This function plays an important role in the identification of any moving average component in the model, something that we will abstract away from, which is why this is a footnote and not a more substantive part of the discussion.

Equation (16) is an example of a quite general result, known as the *Wold decomposition theorem* that states that any covariance-stationary time series can be written as the sum of two time series, one *deterministic* and one *stochastic* (or *indeterministic*).[30] The deterministic component here is $\phi^{r+1}Y_{t-r-1}$, something that can be forecast perfectly given knowledge of its own past (and the parameter ϕ). The stochastic component is a moving average process, also called a *linear process*.

If we take expectations of both sides of (16) we see that

$$\mathrm{E}\left[Y_t\right] = \phi^{r+1}\,\mathrm{E}\left[Y_{t-r-1}\right] + \sum_{j=0}^{r}\phi^j\,\mathrm{E}\left[\epsilon_{t-j}\right] = \phi^{r+1}\,\mathrm{E}\left[Y_{t-r-1}\right],$$

given our assumptions about the ϵ_t. In light of this we observe that ϕ^{r+1} will become diminishingly small if $\mathrm{abs}\left(\phi\right) < 1$. Moreover, if we allow r to diverge to infinity then $\mathrm{E}\left[Y_t\right] = 0$ provided that $\mathrm{abs}\left(\phi\right) < 1$. Our representation then becomes

$$Y_t = \sum_{j=0}^{\infty}\phi^j\epsilon_{t-j}.$$

Another way of obtaining this representation is to re-write (14) as

$$Y_t - \phi Y_{t-1} = \epsilon_t \implies (1-\phi L)Y_t = \epsilon_t \implies Y_t = \frac{\epsilon_t}{(1-\phi L)}.$$

Expanding the term $(1-\phi L)^{-1}$ in series yields

$$\frac{\epsilon_t}{(1-\phi L)} = \sum_{j=0}^{\infty}(\phi L)^j\epsilon_t = \sum_{j=0}^{\infty}\phi^j\epsilon_{t-j},$$

as before.

It is a characteristic of covariance stationary ARMA models that they have in infinite order moving average representation. Turning to the variance of Y_t, given that $\mathrm{Cov}\left[\epsilon_t, \epsilon_{t-s}\right]$ for any integer $s \neq 0$, we see that

$$\gamma(0) = \mathrm{Var}\left[Y_t\right] = \sum_{j=0}^{\infty}\phi^{2j}\,\mathrm{Var}\left[\epsilon_{t-j}\right] = \sigma^2\sum_{j=0}^{\infty}\phi^{2j}.$$

Covariance stationarity in this model therefore requires that there exists some real number M such that

$$\sum_{j=0}^{\infty}\phi^{2j} < M < \infty.$$

We note that if $\phi < 1$ then $\phi^2 < 1$ and so our series is just a convergent geometric progression

$$\sum_{j=0}^{\infty}(\phi^2)^j = \frac{1}{1-\phi^2}$$

so that

$$\gamma(0) = \mathrm{Var}\left[Y_t\right] = \frac{\sigma^2}{1-\phi^2}.$$

[30] This result is named for Herman Wold (1908–1992) a Swedish econometrician who was born in Norway. It first appeared in the first (1938) edition of his book (Wold, 1954).

So we see that covariance stationarity is driven by the value of the parameter ϕ. Indeed, if $\phi \geq 1$ then the Y_t cannot form a stationary time series because (i) the impact of its own infinite past on the mean of the series not only never vanishes but will become dominant, and (ii) the variance of the series diverges to infinity rather than converging to a finite constant.

We can extend these ideas to more general AR processes. Suppose that we have an AR(p) process

$$Y_t = \phi_1 Y_{t-1} + \cdots + \phi_p Y_{t-1} + \epsilon_t.$$

As before, we can re-write this process as

$$\begin{aligned} & Y_t - \phi_1 Y_{t-1} - \cdots - \phi_p Y_{t-1} = \epsilon_t \\ \implies & (1 - \phi_1 L - \cdots - \phi_p L^p) Y_t = \epsilon_t \\ \implies & Y_t = \frac{\epsilon_t}{1 - \phi_1 L - \cdots - \phi_p L^p} \end{aligned}$$

5.3 Non-Stationary Autoregressive Models: The Random Walk

Consider the AR(1) model

$$z_t = \phi z_{t-1} + u_t, \tag{17}$$

where now $\phi = 1$, so that

$$\begin{aligned} z_t &= z_{t-1} + \epsilon_t \\ &= \sum_{s=1}^{t} \epsilon_s \end{aligned} \tag{18}$$

This is known as the *random walk model*. We shall assume that $z_0 = 0$, called an initial condition,[31] and that $\epsilon_t \sim I\,\mathrm{N}(0, \sigma_\epsilon^2)$, $t = 1, ..., T$ where the notation should be read as meaning that the ϵ_t are jointly normally distributed but independent. These are very strong (and largely unnecessary) assumptions made for ease of exposition. Moving forward, we shall be interested in the behaviour of sums of the form

$$\sum_{t=1}^{T} z_t = \sum_{t=1}^{T} \sum_{s=1}^{t} \epsilon_s.$$

As a preliminary we shall define the standardized Wiener process,[32] denoted $W(r)$ (also called Brownian motion).

Definition 4. A Standardized Wiener Process (Brownian Motion).
The standardized Wiener process is a continuous stochastic process on the interval $[0, 1]$, e.g., $W(r) \in C[0, 1]$, if

[31] Initial conditions can be very important to the behaviour of such models but that is too big an area for us to explore here.

[32] Named for the great American mathematician Norbert Wiener (1894–1964). The term Brownian motion is a reference to the botanist Scottish botanist Robert Brown (1773–1858) who had observed pollen jiggling around in water in a random fashion; the apochryphal story is based on tea leaves in a cup of hot water. Albert Einstein's (1879–1955) first contribution to science was to write a paper explaining the cause of this movement, being a result of collisions of the pollen with water molecules, something Brown was unable to do, and to provide a model for this movement based on the Normal distribution.

1. $W(0) = 0$ (a convention of beginning the process at zero).
2. For fixed r, $W(r) = \mathrm{N}(0, r)$.
3. The increments $W(r_i) - W(r_{i-1})$ are stationary and independent with
   (a) $\mathrm{E}\left[W(r_i) - W(r_{i-1})\right] = 0$, $0 \leq r_{i-1} \leq r_i \leq 1$;
   (b) $\mathrm{E}\left[W(r)W(s)\right] = \min(r, s)$;
   (c) and $W(s)$ is independent of $W(r) - W(s)$ $\forall 0 < s < r \leq 1$. □

The next step is to define

$$X_T(r) = [\sigma_\epsilon \sqrt{T}]^{-1} S_{\lfloor Tr \rfloor} \quad 0 \leq r \leq 1$$

where $S_0 = 0$ and $S_i = \sum_{s=1}^{i} \epsilon_S$ $\forall i = 1, ..., T$ and $\lfloor Tr \rfloor$ denotes the integer part of Tr, eg $\lfloor Tr \rfloor = 0$ $\forall r < \frac{1}{T}$; $\lfloor Tr \rfloor = 1$, $\forall \frac{1}{T} \leq r < \frac{2}{T}$; etc.

Observe that

$$\begin{aligned}
S_{\lfloor Tr \rfloor} &= S_0 = 0 \quad \forall r < \frac{1}{T} \\
&= S_1 = \epsilon \quad \forall \frac{1}{r} \leq r < \frac{2}{T} \\
&\vdots \\
&= S_{i-1} = \sum_{s=1}^{i-1} \epsilon_s \quad \forall \frac{i-1}{T} \leq r < \frac{i}{T} \quad i = 1, ..., T \\
&= S_T = \sum_{s=1}^{T} \epsilon_s \quad r = 1.
\end{aligned}$$

We wish to establish that $X_T(r) \implies W(r)$ where $\implies$ denotes weak convergence (the equivalent of convergence in distribution of random variables when working with stochastic processes). To begin,

$$\begin{aligned}
\mathrm{E}\left[X_t(r)\right] &= \left[\sigma_\epsilon \sqrt{T}\right]^{-1} \mathrm{E}\left[S_{\lfloor Tr \rfloor}\right] \\
&= \left[\sigma_\epsilon \sqrt{T}\right]^{-1} \begin{cases} \mathrm{E}\left[0\right], & r = 0, \\ \mathrm{E}\left[\sum_{s=1}^{\lfloor Tr \rfloor} \epsilon_s\right], & 0 < r \leq 1, \end{cases}
\end{aligned}$$

as $\mathrm{E}\left[\sum_{s=1}^{\lfloor Tr \rfloor} \epsilon_s\right] = \sum_{s=1}^{\lfloor Tr \rfloor} \mathrm{E}\left[\epsilon_s\right] = 0$ as $\mathrm{E}\left[\epsilon_s\right] = 0$ $\forall s$ (by assumption). Next,

$$\begin{aligned}
\mathrm{E}\left[X_T^2(r)\right] &= \left[\sigma_\epsilon^2 T\right]^{-1} \mathrm{E}\left[S_{\lfloor Tr \rfloor}^2\right] \\
&= \begin{cases} 0, & \text{if } r < \frac{1}{T}, \\ \mathrm{E}\left[\left[\sigma_\epsilon^2 T\right]^{-1} \left[\sum_{s=1}^{\lfloor Tr \rfloor} \epsilon_s^2 + 2 \sum_{i \neq j}^{\lfloor Tr \rfloor} \epsilon_i \epsilon_j\right]\right], & \text{otherwise}, \end{cases} \\
&= \begin{cases} \left[\sigma_\epsilon^2 T\right]^{-1} (j-1)\sigma_\epsilon^2, & \text{if } \frac{j-1}{T} \leq r < \frac{j}{T} \; j = 1, ..., T \\ \left[\sigma_\epsilon^2 T\right]^{-1} T \sigma_\epsilon^2, & \text{if } r = 1, \end{cases} \\
&= \begin{cases} \frac{j-1}{T}, & \text{if } \frac{j-1}{T} \leq r < \frac{j}{T} \; j = 1, ..., T, \\ 1, & \text{if } r = 1. \end{cases}
\end{aligned}$$

Note that as $T \to \infty$, $\frac{j}{T} - \frac{j-1}{T} = \frac{1}{T} \to 0 \implies \frac{j-1}{T} \to r$ as $T \to \infty$, i.e., r lies in an interval of diminishing length and must approach the boundary of the interval as $T \to \infty$. Consequently,

$$X_T(r) = \left[\sigma_\epsilon \sqrt{T}\right]^{-1} \left(\epsilon_1 + ... + \epsilon_{\lfloor Tr \rfloor}\right) \sim \mathrm{N}\left(0, \frac{j-1}{T}\right),$$

as $\epsilon_1 + ... + \epsilon_{\lfloor Tr \rfloor}$ is just a sum of independent normal random variables

$$\mathrm{N}\left(0, \frac{j-1}{T}\right) \to \mathrm{N}\left(0, r\right),$$

as $T \to \infty$. Now,

$$X_T\left(r + \frac{1}{T}\right) - X_T\left(r\right) = \left[\sigma_\epsilon \sqrt{T}\right]^{-1} \epsilon_{\left[T\left(r + \frac{1}{T}\right)\right]} \sim I\,\mathrm{N}\left(0, \frac{1}{T}\right) \ \forall r.$$

So,

$$\mathrm{E}\left[X_t\left(r + \frac{1}{T}\right) - X_T\left(r\right)\right] = 0.$$

Next,

$$\begin{aligned}
\mathrm{E}\left[X_T\left(r\right) X_T\left(s\right)\right] &= \mathrm{E}\left[\sigma_\epsilon^2 T\right]^{-1} \sum_{i=1}^{\lfloor Tr \rfloor} \sum_{i=1}^{\lfloor Ts \rfloor} \epsilon_i \epsilon_j \\
&= \left[\sigma_\epsilon^2 T\right]^{-1} \sum_{k=1}^{\lfloor Tp \rfloor} \mathrm{E}\left[\epsilon_k^2\right], \qquad \text{where } \underbrace{p = \min\left(r, s\right)}_{\text{Follows from } \mathrm{E}[\epsilon_i \epsilon_j] = 0 \ \forall i \neq j}, \\
&= \frac{\lfloor Tp \rfloor \sigma_\epsilon^2}{T \sigma_\epsilon^2} = p.
\end{aligned}$$

Finally,

$$\begin{aligned}
\mathrm{E}\left[X_T(s)\left(X_T(r) - X_T(s)\right)\right] &= \mathrm{E}\left[X_T(s) X_T(r)\right] - \mathrm{E}\left[X_T^2(s)\right] \\
&= s - s \\
&= 0 \quad \forall 0 < s \leq 1
\end{aligned}$$

This implies that $X_T(s)$ and $X_T(r) - X_T(s)$ are independent as they are jointly normally distributed normal variables with zero covariance. Given that $\epsilon_s \sim I\,\mathrm{N}(0, \sigma_\epsilon^2)$, $s = 1, ..., T$ it follows that $X_T(r) \to W(r)$ as $T \to \infty$.

Note A Wiener process is continuous and $X_T(r)$ is discrete. As $T \to \infty$,

$$\lim_{T \to \infty} \left[\left(r + \frac{1}{T}\right) - r\right] = 0,$$

so the steps converge to intervals of zero length and $X_T(r)$ converges to something that is continuous. A consequence of what has been established is that $W(1) \equiv \mathrm{N}(0, 1)$, which is implied by standard CLT's for random variables. For example,

$$X_T(1) = \left[\sigma_\epsilon \sqrt{T}\right]^{-1} \sum_{s=1}^{T} \epsilon_s \xrightarrow{d} \mathrm{N}(0, 1).$$

5.4 An Aside on the Continuous Mapping Theorem

The Continuous Mapping Theorem (CMT) essentially says that if $g(V)$ is a continuous function of V with probability one and if

$$V_T \xrightarrow{d} V.$$

Then

$$g(V_T) \xrightarrow{d} g(V).$$

In essence we are saying that, provided $g(\cdot)$ is continuous

$$\operatorname*{plim}_{T\to\infty} g(V_T) = g\left(\operatorname*{plim}_{T\to\infty} V_T\right)$$

which is a powerful result and is behind Slutsky's Theorem, etc. We can extend this result to give the so-called Invariance Principle or Functional Central Limit Theorem.

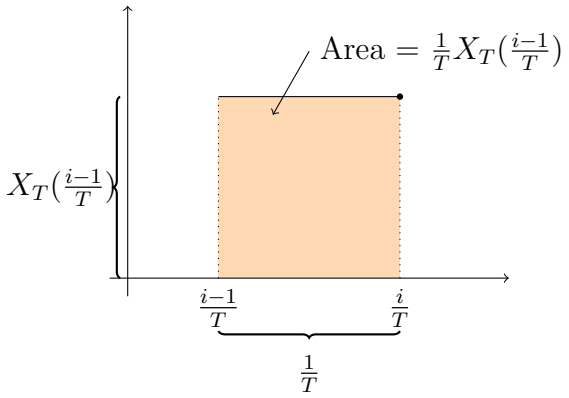
Theorem 17. Invariance Principle *If* $X_T(r) \implies W(r)$ *and* h *is any continuous functional on* $C[0,1]$ *then* $h(X_T(r)) \implies h(W(r))$.

Comment This extension to the CMT allows us to map discrete quantities, $X_T(r)$, to functions of stochastic processes.

As $X_T(r)$ is a step function changing only at integer values of Tr it follows that

$$\int_{\frac{i-1}{T}}^{\frac{i}{T}} X_T(r)\,\mathrm{d}r = \frac{1}{T} X_T\left(\frac{i-1}{T}\right),\ i = 1, ..., T$$

Note Within a given interval $\left[\frac{i-1}{T}, \frac{i}{T}\right)$, $X_T(r)$ is not a function of r as it is not changing with r as r moves through the interval. Graphically



This implies that

$$\frac{1}{T}\sum_{t=1}^{T} X_T\left(\frac{t-1}{T}\right) = \sum_{t=1}^{T}\int_{\frac{t-1}{T}}^{\frac{t}{T}} X_T(r)\,\mathrm{d}r$$

but

$$\sum_{t=1}^{T}\int_{\frac{t-1}{T}}^{\frac{t}{T}} \equiv \int_0^1.$$

Hence

$$\frac{1}{T}\sum_{t=1}^{T} X_T\left(\frac{t-1}{T}\right) = \int_0^1 X_T(r)\,\mathrm{d}r \implies \int_0^1 W(r)\,\mathrm{d}r \text{ as } T\to\infty$$

An alternative of writing this is

$$\frac{1}{T}\left[\sigma_\epsilon\sqrt{T}\right]^{-1} S_{\lfloor Tr\rfloor} = [\sigma_\epsilon T^{\frac{3}{2}}]^{-1} S_{\lfloor Tr\rfloor} \implies \int_0^1 W(r)\,\mathrm{d}r$$

$$\text{or } T^{-\frac{3}{2}} S_{\lfloor Tr\rfloor} \implies \sigma_\epsilon \int_0^1 W(r)\,\mathrm{d}r$$

Note that there is a variety of ways in which the various assumptions can be relaxed. As an indication, so long as observations are independent the sum of them will converge to normality and so normality of the ϵ's need not be assumed. Neither need constancy of variance or even independence. There is, however, no such thing as a free lunch. The more general the model the more complicated the analysis.

The preceding allows us to establish the following:

$$\mathrm{E}\left[\int_0^1 W(r)\,\mathrm{d}r\right] = 0$$

$$\mathrm{Var}\left[\int_0^1 W(r)\,\mathrm{d}r\right] = \frac{1}{3}$$

Together these results imply $\int_0^1 W(r)\,\mathrm{d}r \equiv \mathrm{N}(0, \frac{1}{3})$

Proof Consider

$$\int_0^1 W(r)\,\mathrm{d}r = \lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T} W\left(\frac{t}{T}\right)$$

As $W(\frac{t}{T})$ is normal then so is the sum. Similarly, each term in the sum has zero mean and hence so does the sum.

$$\begin{aligned}
\mathrm{Var}\left[\int_0^1 W(r)\,\mathrm{d}r\right] &= \mathrm{E}\left[\left(\int_0^1 W(r)\,\mathrm{d}r\right)^2\right] \\
&= \mathrm{E}\left[\int_0^1\int_0^1 W(r)W(s)\,\mathrm{d}s\,\mathrm{d}r\right] \\
&= \mathrm{E}\left[\int_0^1\int_0^r W(r)W(s)\,\mathrm{d}s\,\mathrm{d}r\right] + \mathrm{E}\left[\int_0^1\int_r^1 W(r)W(s)\,\mathrm{d}s\,\mathrm{d}r\right] \\
&= \underbrace{\int_0^1\int_0^r}_{s\le r}\mathrm{E}\left[\underbrace{W(r)W(s)}_{\min(r,s)=s}\right]\mathrm{d}s\,\mathrm{d}r + \underbrace{\int_0^1\int_r^1}_{s\ge r}\mathrm{E}\left[\underbrace{W(r)W(s)}_{\min(r,s)=r}\right]\mathrm{d}s\,\mathrm{d}r \\
&= \int_0^1\int_0^r s\,\mathrm{d}s\,\mathrm{d}r + \int_0^1\int_r^1 r\,\mathrm{d}s\,\mathrm{d}r \quad \text{as } \mathrm{E}\left[W(r)W(s)\right] = \min(r,s) \\
&= \int_0^1 \frac{s^2}{2}\bigg|_0^r \mathrm{d}r + \int_0^1 rs\bigg|_r^1 \mathrm{d}r \\
&= \int_0^1 \left[\frac{r^2}{2} + r - r^2\right]\mathrm{d}r = \int_0^1 \left[r - \frac{r^2}{2}\right]\mathrm{d}r
\end{aligned}$$

$$= \left[\frac{r^2}{2} - \frac{r^3}{6}\right]_0^1 = \frac{1}{2} - \frac{1}{6} = \frac{1}{3} \text{ as required.}$$

5.5 OLS Estimation in the Random Walk Model

Here we wish to explore the behaviour of the OLS estimator in the random walk model. The OLS estimator of the coefficient on z_{t-1} in an AR(1) model (17) is

$$\hat{\phi} = \frac{\sum_{t=1}^T z_{t-1} z_t}{\sum_{t=1}^T z_{t-1}^2} = \phi + \frac{\sum_{t=1}^T z_{t-1} u_t}{\sum_{t=1}^T z_{t-1}^2} \tag{19}$$

We can (but won't) show that if $|\phi| < 1$, so that the model is stationary, then

$$\sqrt{T}(\hat{\phi} - \phi) \xrightarrow{d} \mathrm{N}(0, 1 - \phi^2). \tag{20}$$

If we think about the behaviour as $\phi \to 1^-$, that is as ϕ approaches unity from below, then (20) suggests that

$$\sqrt{T}(\hat{\phi} - \phi) \xrightarrow{d} \mathrm{N}(0, 0)$$

That is, in the case of a random walk, the limiting distribution of the OLS estimator $\hat{\phi}$ is becoming degenerate (zero variance) but centred at zero so that $\hat{\phi}$ is consistent for ϕ.[33] This is a good thing! However, it is not helpful for inference. In particular, it does not lend itself to either hypothesis testing and confidence intervals. However, it transpires that (20) provides a poor approximation to the actual behaviour of $\hat{\phi}$ in the random walk model, it is unreasonably optimistic, and so let's try to do better.

To begin, from (18) we see that

$$z_t = \sum_{s=1}^t u_s \sim \mathrm{N}(0, \sigma_u^2 t) \qquad \text{because } z_0 = 0 \text{ by assumption.} \tag{21}$$

That is,

$$\begin{aligned} z_1 &= z_0 + u_1 = u_1 \quad (\text{because } z_0 = 0 \text{ by assumption}) \\ z_2 &= z_1 + u_2 = u_1 + u_2 \\ &\vdots \\ z_T &= u_1 + \ldots + u_T \end{aligned}$$

Moreover,

$$\begin{aligned} z_t^2 &= (z_{t-1} + u_t)^2 = z_{t-1}^2 + 2 z_{t-1} u_t + u_t^2 \\ \implies z_{t-1} u_t &= \frac{1}{2}\left[z_t^2 - z_{t-1}^2 - u_t^2\right] \end{aligned}$$

Hence, the numerator on the right-hand side of (19) can be written

$$\sum_{t=1}^T z_{t-1} u_t = \frac{1}{2}\left[z_T^2 - z_{T-1}^2 - u_T^2\right] + \frac{1}{2}\left[z_{T-1}^2 - z_{T-2}^2 - u_{T-1}^2\right] + \ldots + \left[z_1^2 - z_0^2 - u_1^2\right]$$

[33] In fact, in the random walk model $\hat{\phi}$ is said to be *super-consistent* because the rate of convergence to itś probability limit is actually faster than in the stationary case, but that is a story for another time.

$$= \frac{1}{2} z_T^2 - \frac{1}{2} \sum_{t=1}^{T} u_t^2$$

Re-scaling, we obtain

$$\frac{1}{\sigma_u^2 T} \sum_{t=1}^{T} z_{t-1} u_t = \frac{1}{2} \left(\frac{z_T}{\sigma_u \sqrt{T}} \right)^2 - \frac{1}{2\sigma_u^2 T} \sum_{t=1}^{T} u_t^2.$$

Observe that our normality assumption implies, from (21), that

$$\frac{z_T}{\sigma_u \sqrt{T}} \sim \mathrm{N}\left(0, \frac{\sigma_u^2 T}{(\sigma_u \sqrt{T})^2}\right) = \mathrm{N}(0,1).$$

Moreover, writing $S_T = \frac{1}{T} \sum_{t=1}^{T} u_z^2$, with $\mathrm{E}\left[u_t^2\right] = \sigma_u^2$ and u_t iid (which implies u_t^2 iid), the Law of Large Numbers implies that

$$S_T \xrightarrow{p} \sigma_u^2$$

That is,

$$\lim_{T\to\infty} S_T = \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} u_t^2 = \mathrm{E}\left[u_t^2\right] = \sigma_u^2.$$

Gathering these results, we see that the scaled numerator of (19) converges to a random variable as follows

$$\frac{1}{\sigma_u^2 T} \sum_{t=1}^{T} z_{t-1} u_t \xrightarrow{d} \frac{1}{2}[(\mathrm{N}(0,1))^2 - 1] = \frac{1}{2}\left(\chi_1^2 - 1\right)$$

Next consider the denominator from (19): $\left(\sum_{t=1}^{T} z_{t-1}^2\right)$. Recall from (21) that

$$z_{t-1} \sim \mathrm{N}(0, \sigma_u^2 (t-1))$$

and so

$$\frac{z_{t-1}}{\sqrt{\sigma_u^2 (t-1)}} \sim \mathrm{N}(0,1) \implies \left(\frac{z_{t-1}}{\sqrt{\sigma_u^2 (t-1)}} \right)^2 \sim \chi_1^2.$$

Recall from the properties of the chi-squared distribution that if $Q \sim \chi_\nu^2$ then $\mathrm{E}[Q] = \nu$. Hence,

$$\mathrm{E}\left[\frac{z_{t-1}^2}{\sigma_u^2 (t-1)} \right] = 1 \implies \mathrm{E}\left[z_{t-1}^2\right] = \sigma_u^2 (t-1).$$

Consequently,[34]

$$\mathrm{E}\left[\sum_{t=1}^{T} z_{t-1}^2 \right] = \sum_{t=1}^{T} \mathrm{E}\left[z_{t-1}^2\right] = \sigma_u^2 \sum_{t=1}^{T} (t-1) = \frac{\sigma_u^2 (T-1) T}{2} = O(T^2).$$

[34] Observe first that $\sum_{t=1}^{T} (t-1) = 0 + 1 + \cdots + (T-1) = \sum_{t=1}^{T-1} t$. Second $\sum_{n=1}^{N} n = N(N+1)/2$ is a well-known result. To see why it must be true consider the problem of counting the number of terms on or below the main diagonal of an $N \times N$ rectangular array. For example, look at the 3×3 array below.

$$\begin{array}{ccc} a & b & c \\ d & e & f \\ g & h & i \end{array}$$

The terms to be counted are a, d, e, g, h, i. That is, there are 6 such terms. Observe that we have one

So we can see why the limiting distribution of $\sqrt{T}(\hat{\phi}-\phi)$ is degenerate when $\phi=1$, the denominator is of order T^2 and so scaling by $T^{\frac{1}{2}}$ is insufficient to stop it vanishing. We need to scale by more. Specifically

$$T(\hat{\phi}-1)=\left(\frac{1}{\sigma_u^2 T}\sum_{t=1}^{T} z_{t-1}u_t\right)\left(\frac{1}{\sigma_u^2 T^2}\sum_{t=1}^{T} z_{t-1}^2\right)^{-1}$$

Observe that

$$\begin{aligned}
T^{-2}\sum_{t=1}^{T} z_{t-1}^2 &= \sigma_u^2 T^{-1}\sum_{t=1}^{T}\left\{\left[\sigma_u\sqrt{T}\right]^{-1} z_{t-1}\right\}^2 \\
&= \sigma_u^2 T^{-1}\sum_{t=1}^{T}\left\{X_T(r)\right\}^2, \text{ where } X_T(r)=\frac{z_{t-1}}{\sigma_u\sqrt{T}} \\
&\xrightarrow{d} \sigma_u^2\int_0^1 [W(r)]^2\,\mathrm{d}r
\end{aligned}$$

because $X_T(r)\to W(r)$ as $T\to\infty$.

Finally, recalling that $\mathrm{N}(0,1)\equiv W(1)$, it follows that $\chi_1^2\equiv[W(1)]^2$ and so we see that

$$T(\hat{\phi}_T-1)\xrightarrow{d}\frac{\frac{1}{2}\left\{[W(1)]^2-1\right\}}{\int_0^1[W(r)]^2\,\mathrm{d}r} \tag{22}$$

So we see that in the random walk model the OLS estimator has a non-standard distribution. We can't learn too much about such results by inspection and so we simulate. As an aside, even thought this non-standard distribution is strictly only applicable in the case of a random walk, you might find that (22) provides a better approximation to the true sampling distribution of $\hat{\phi}$ than does (20) for $|\phi|$ close to unity.

The obvious question is 'How do we know which limiting distribution to use?' (Clearly we wouldn't bother estimating if we knew the true value of ϕ.) This choice is typically made on the basis of an hypothesis test, where the hypotheses under test are

$$\begin{aligned}
&H_0:\phi=1 && \text{(non-stationary model)}\\
&H_1:|\phi|<1. && \text{(stationary model)}
\end{aligned}$$

We can extend the analysis to consider a t-statistic which is the obvious test of these hypothesis.

---

term from the first row, two from the second and three from the third. Continuing this pattern, in the case of an $N\times N$ array, we need to sum $1+2+3+\cdots+N=\sum_{n=1}^{N} n$, which is the problem we want to solve. Okay, from here on in it is a pretty simple thing to do. We know that there are N^2, in an $N\times N$ array. So $N^2/2$ gives us half the elements, but this only counts half of the elements on the main diagonal, so we need to add another $N/2$ elements to our count. Thus,

$$\sum_{n=1}^{N} n=\frac{N^2}{2}+\frac{N}{2}=\frac{N(N+1)}{2}.$$

To check our result, setting $N=3$, we find that there are 6 terms on or below the main diagonal of the array, which is (of course) the expected answer. Returning to our problem, we need to set $N=T-1$, which yields $(T-1)T/2$, as stated.

Write

$$t = \frac{\hat{\phi} - 1}{\hat{\sigma}_{\hat{\phi}}} = \frac{\hat{\phi} - 1}{\sqrt{s^2 / \sum_{t=1}^{T} z_{t-1}^2}}$$

where

$$s^2 = \frac{1}{T-1} \sum_{t=1}^{T} (z_t - \hat{\phi} z_{t-1})^2$$

is the usual OLS estimate of the disturbance variance. Because $\hat{\phi} \xrightarrow{a.s.} \phi = 1$, under H_0

$$s^2 = \frac{1}{T-1} \sum_{t=1}^{T} (z_t - \hat{\phi} z_{t-1})^2 \xrightarrow{a.s.} \lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} u_t^2 = \mathrm{E}\left[u_t^2\right] = \sigma_u^2$$

That is, $s^2 \xrightarrow{a.s.} \sigma_u^2$.

Using similar analysis to that above we can show that

$$t \xrightarrow{d} \frac{\frac{1}{2}\sigma_u^2 \left\{[W(1)]^2 - 1\right\}}{\left[\sigma_u^2 \int_0^1 [W(r)]^2 \, \mathrm{d}r\right]^{\frac{1}{2}} (\sigma_u^2)^{\frac{1}{2}}} = \frac{\frac{1}{2}\left\{[W(1)]^2 - 1\right\}}{\left[\int_0^1 [W(r)]^2 \, \mathrm{d}r\right]^{\frac{1}{2}}}$$

This is the so-called Dickey-Fuller distribution which is tabulated in a variety of places.

Bibliography

Amemiya, T. (1985). *Advanced Econometrics*. Harvard University Press, Cambridge, Massachusetts. 4, 23

Chung, K. L. (1968). *A Course in Probability Theory*. Harcourt, New York. 24

Cramér, H. (1946). *Mathematical Methods of Statistics*. Princeton University Press, Princeton, New Jersey. 4, 23

Gallant, A. R. (1997). *An Introduction to Econometric Theory: Measure-Theoretic Probability and Statistics with Applications to Economics*. Princeton University Press, Princeton, New Jersey. 4

Hoeffding, W. and H. Robbins (1948). The central limit theorem for dependent variables. *Duke Mathematical Journal* 15(3), 773–780. 19

Malinvaud, E. (1970). *Statistical Methods of Econometrics*. North Holland Publishing Company, Amsterdam, second revised edition. 19

Newey, W. K. and D. McFadden (1994). Large sample estimation and hypothesis testing. In *Handbook of Econometrics, Volume IV*, R. F. Engle and D. L. McFadden, editors, chapter 36, 2111–2245, Elsevier Science B.V., Amsterdam. 4

Rao, C. R. (1973). *Linear Statistical Inference and Its Applications*. John Wiley & Sons, Inc., New York, second edition. 4, 15

Schmidt, P. (1976). *Econometrics*. Marcel Dekker, New York. 4

Schönfeld, P. (1971). A useful central limit theorem for m-dependent variables. *Metrika* 17(1), 116–128, ISSN 1435-926X, doi:10.1007/BF02613816. 20

Theil, H. (1971). *Principles of Econometrics*. John Wiley and Sons, New York. 4

Wald, A. (1949). Note on the consistency of the maximum likelihood estimate. *Annals of Mathematical Statistics* 20(4), 595–601. 4

Wold, H. O. A. (1954). *A Study in the Analysis of Stationary Time Series*. Almqvist and Wiksell Book Co., Uppsala, second revised edition. 34

A Sequences and Series

A.1 Sequences, Series and Partial Sums

Our analysis of asymptotics requires the evaluation of partial sums of terms in a sequence, sometimes a geometric sequence. Before dealing with geometric sequences specifically, we shall introduce some definitions that apply to sequences generally.

An Infinite Sequence.
An infinite sequence is a real function whose domain is the set of positive integers. The values taken by the function are called the terms of the sequence.

It is usual to omit the word 'infinite' and refer simply to a sequence. We typically denote a sequence by writing out some of its terms, e.g. $x_1, x_2, x_3, \ldots$. The practical importance of this definition is that the subscripts have meaning, they indicate order. Hence x_1 comes before x_2, and x_2 comes between x_1 and x_3, etc.

Infinite Series.
If $x_1, x_2, x_3, \ldots$ denotes a sequence then an infinite series is the value taken by the sum of all of the terms in the sequence.

We usually omit the word 'infinite' and speak simply of *series*.

Partial Sums.
If $x_1, x_2, x_3, \ldots$ denotes a sequence then a partial sum is the sum of a finite number of terms from the sequence.

The partial sum of the first n terms in a sequence is often denoted by S_n. Thus,

$$S_n = x_1 + x_2 + \ldots + x_n.$$

Using this notation, the infinite series is then $S_\infty = x_1 + x_2 + \ldots$

Geometric Sequence.
A geometric sequence has a typical term of the form $f(i) = ax^{i-1}$ $(a \neq 0)$, $i = 1, 2, 3, \ldots$.

Let us illustrate some of these ideas by examining the partial sum of a geometric sequence. It is clear that if $x = 1$ then $S_n = a + a + \ldots + a = na$. Similarly, if $x = -1$ then

$$S_n = a - a + a - \ldots \pm a = \begin{cases} a, & \text{if } n \text{ is odd,} \\ 0, & \text{if } n \text{ is even} \end{cases}$$

However, if $|x| \neq 1$ evaluating the partial sum is trickier. Observe that

$$\begin{array}{rcccccccccc} S_n - xS_n = & a & + & ax & + & \ldots & + & ax^{n-1} & & \\ & & - & ax & - & \ldots & - & ax^{n-1} & - & ax^n \\ \hline = a - ax^n. & & & & & & & & & \end{array}$$

Simple rearrangement yields $S_n = a(1 - x^n)/(1 - x)$. Combining these results yields

$$S_n = \begin{cases} \dfrac{a(1-x^n)}{1-x}, & |x| \neq 1, \\ na, & x = 1, \\ a, & x = -1 \text{ and } n \text{ is odd}, \\ 0, & x = -1 \text{ and } n \text{ is even} \end{cases} \tag{A.1}$$

The various cases for the infinite series S_∞ can be deduced by considering what happens as n becomes large. If $|x| < 1$, x^n approaches zero as n becomes large, and $S_\infty = a/(1-x)$. If $x \geq 1$ then S_∞ will diverge, with its sign being that of a, denoted $\operatorname{sgn}(a)$. If $x \leq -1$ then S_n will oscillate as the value of n increases. In the case of $x = -1$ it will oscillate between 0 and a and so will be bounded, whereas if $x < -1$ the sum will be unbounded. In summary,

$$S_\infty = \begin{cases} \dfrac{a}{1-x}, & \text{if } |x| < 1, \\ \text{oscillates between 0 and } a, & \text{if } x = -1 \\ \text{diverges}, & \text{otherwise.} \end{cases}$$

A.2 Power Series and Taylor's Theorem

A.2.1 Power Series

Sometimes it is convenient to find alternative representations for functions, representations that may be easier to work with. One such representation expresses the function, evaluated at some point x, as a power series, which is an infinite series of the form

$$f(x) = a_0 + a_1(x-c) + a_2(x-c)^2 + a_3(x-c)^3 + \cdots \tag{A.2}$$

where a_i represents the coefficient of the ith term, c is a constant, and x varies around c (for this reason one sometimes speaks of the series as being centred at c). The coefficients a_i are not functions of x.

A power series will converge for some values of the variable x (at least for $x = c$) and may diverge for others. There is always a number r, with $0 \leq r < \infty$, such that the series converges whenever $|x-c| < r$ and diverges whenever $|x-c| > r$. The number r is called the radius of convergence of the power series. A fast way to compute r is

$$r = \lim_{i \to \infty} \left| \frac{a_{i+1}}{a_i} \right|$$

if this limit exists. The series converges absolutely for $|x-c| < r$ and converges uniformly on every compact subset of $\{x : |x-c| < r\}$. For $|x-c| = r$, we cannot make any general statement on whether the series converges or diverges.

A.2.2 Taylor Series

The definition of power series provided in Section A.2.1 is all very well at a theoretical level but the great practical question is how does one find the coefficients a_i. A solution to this question for differentiable functions leads to Taylor's Theorem. Specifically, Taylor showed that, provided the derivatives existed,

$$a_0 = f(c),$$

and

$$a_i = \frac{1}{i!} \left. \frac{d^i f(x)}{dx^i} \right|_{x=c}, \quad i = 1, 2, \ldots$$

The problem with this infinite series representation is simply that it involves an infinite series. Hence, it is only applicable to infinitely differentiable functions f and, as one can

never sum an infinite number of terms in finite time, there are only certain special cases where the series can be evaluated exactly; one such example is the geometric series

$$\sum_{i=0}^{\infty} sx^i = \frac{s}{1-x}, \quad |x| < 1.$$

These observations led to the notion of breaking the power series into two components, one the sum of a finite number of terms and the other a remainder. Thus, we might express f evaluated at x as a polynomial of order S plus a remainder, denoted R_S,

$$f(x) = \sum_{i=0}^{S} a_i (x-c)^i + R_S$$

where

$$R_S = a_{S+1}(x-c)^{S+1} + a_{S+2}(x-c)^{S+2} + a_{S+3}(x-c)^{S+3} + \cdots$$

The advantage of such a representation is that n need be no large than the number of well-defined derivatives at c and can be kept small enough to make the sum manageable. The problem of course is that the remainder term R_S must either be evaluated or be demonstrated to be irrelevant. Whilst in the form of an infinite series, as it currently is, that task is every bit a problematic as dealing with the original power series, there exist a number of different representations for R_S specifically designed to address this problem. The form of the remainder that we shall use is known as the Lagrange form, which states that

$$R_S = \frac{1}{(S+1)!} \left. \frac{d^{(S+1)} f(x)}{dx^{(S+1)}} \right|_{x=\xi}, \quad \xi = \lambda x + (1-\lambda)c, \quad 0 < \lambda < 1.$$

This definition for R_S introduces the point ξ which lies somewhere on the line segment joining x and c. The result of Lagrange does not tell us the value of ξ, merely that it exists in the stated interval.

Theorem 18 (Univariate Taylor's Theorem with Lagrange Form of Remainder)**.** *Let the function f and its first $S+1$ derivatives be continuous on an interval containing the line segment between the points x and c. Then,*

$$f(x) = f(c) + \frac{f^{(1)}(c)}{1!}(x-c) + \frac{f^{(2)}(c)}{2!}(x-c)^2 + \ldots + \frac{f^{(S)}(c)}{S!}(x-c)^n + R_S(x), \quad \text{(A.3)}$$

where there exists a value ξ such that

$$R_S(x) = \frac{f^{(S+1)}(\xi)}{(S+1)!}(x-c)^{S+1}, \quad \xi = \lambda x + (1-\lambda)c, \quad 0 < \lambda < 1,$$

and we have used the notation

$$f^{(S)}(c) = \left. \frac{d^S f(x)}{dx^S} \right|_{x=c}.$$

It is worth noting that (A.3) is not an approximation, it is an exact result. Further, you should remember that the coefficients of this polynomial are not functions of x. That is, $f^{(s)}(c) \equiv f^{(s)}(x)\big|_{x=a}$ is not a function of x, but rather is a function of c.

We will frequently use the following multivariate version of this result.

Theorem 19 (Multivariate Taylor's Theorem with Lagrange Form of Remainder)**.** *Consider the vector-valued function $f : \mathcal{D} \to \mathcal{D}$, where the domain of f is convex, so that the line segment joining any two points in $\mathcal{D}$ lies completely in $\mathcal{D}$. Let all partial derivatives f of order $S+1$ be continuous for all points in $\mathcal{D}$. Then,*

$$f(x) = f(c) + df(x)|_{x=c} + \frac{1}{2!}d^2 f(x)\bigg|_{x=c} + \cdots + \frac{1}{S!}d^S f(x)\bigg|_{x=c} + R_S, \tag{A.4}$$

where

$$R_S = \frac{d^{S+1}f(x)}{(S+1)!}\bigg|_{x=\xi}, \quad \xi = \lambda x + (1-\lambda)c, \quad 0 < \lambda < 1, \tag{A.5}$$

which implies that ξ is a point on the line segment joining the points c and x, and the terms $d^p f$ are defined to be

$$d^p f(x) = \left[(x-c)'\frac{\partial}{\partial x}\right]^p f(x), \quad p = 1, \ldots, S+1, \tag{A.6}$$

with $\frac{\partial}{\partial x} = \left[\frac{\partial}{\partial x_1}, \ldots, \frac{\partial}{\partial x_p}\right]'$ for any p-vector $x = [x_1, \ldots, x_p]'$. □

B Inequalities

B.1 The Triangle Inequality

It can be shown that the length of a given side of a triangle must be less than or equal to the sum of the lengths of the other two sides, with equality when the angle formed between the other two sides is 180 degrees, but greater than or equal to the difference between these two lengths, with equality when the angle formed by the other two sides is zero degrees.

In its simplest form the triangle inequality states that, for any two real numbers x and y

$$|x + y| \leq |x| + |y|,$$

which clearly reflects the first part of the preceding statement. The second part of this statement is reflected in the inverse triangle inequality

$$\big||x| - |y|\big| \leq |x + y|.$$

We are most likely to encounter the triangle inequality in the form

$$\sqrt{(x+y)'(x+y)} \leq \sqrt{x'x} + \sqrt{y'y},$$

where now x and y are both k-vectors, and the inverse triangle equality in the form

$$\left|\sqrt{x'x} - \sqrt{y'y}\,\right| \leq \sqrt{(x+y)'(x+y)}.$$

B.2 The Cauchy-Schwarz Inequality

In econometrics the Cauchy-Schwarz inequality is frequently encountered in both summation and integral forms. Specifically, if x and y are both k-vectors,

$$|x'y| = \sqrt{x'x}\,\sqrt{y'y}\,,$$

also sometimes written $(x'y)^2 \leq (x'x)(y'y)$, which is often more convenient to work with. We see that, whereas the triangle inequality is concerned with sums, the Cauchy-Schwarz inequality is concerned with products. The integral form of Cauchy-Schwarz is, in its simplest form,

$$\left|\int f(x)g(x)\,dx\right|^2 \leq \int |f(x)|^2\,dx \cdot \int |g(x)|^2\,dx.$$

More useful to us is the following generalization for random k-vectors x and y:

$$|\mathrm{E}\,[x'y]|^2 \leq \mathrm{E}\,[x'x]\,\mathrm{E}\,[y'y]$$

which is the form in which we can establish results such as

$$(\mathrm{Cov}\,[X,Y])^2 \leq \mathrm{Var}\,[X]\,\mathrm{Var}\,[Y]\,,$$

from which we infer that correlations lie in the interval $[-1,1]$. A generalized Cauchy-Schwarz inequality that is sometimes useful is

$$\mathrm{E}\,[yy'] \geq \mathrm{E}\,[yx']\,(\mathrm{E}\,[xx'])^{-1}\,\mathrm{E}\,[xy']\,,$$

for random x and y.

B.3 The Covariance Inequality

Theorem 20 (Covariance Inequality)**.** *Let X be any random variable and $g(x)$ and $h(x)$ be any random variable such that $\mathrm{E}\,[g(X)]$, $\mathrm{E}\,[h(X)]$, and $\mathrm{E}\,[g(X)h(X)]$ exist.*

1. *If $g(x)$ is a non-decreasing function and $h(x)$ a non-increasing function, then*

$$\mathrm{E}\,[g(X)h(X)] \leq (\mathrm{E}\,[g(X)])\,(\mathrm{E}\,[h(X)])\,.$$

2. *If $g(x)$ and $h(x)$ are either both non-decreasing or both non-increasing, then*

$$\mathrm{E}\,[g(X)h(X)] \geq (\mathrm{E}\,[g(X)])\,(\mathrm{E}\,[h(X)])\,.$$

The intuition here is that in Case 1 the functions g and h are negatively correlated, whereas in Case 2 there is positive correlation, which is what is implied by the directions of the inequalities. The usefulness of the inequality is that it sometimes makes it possible to bound an expectation without using higher-order moments. This is clearly closely related to Hölder's inequality (Section B.5).

B.4 The Chebyshev and Markov Inequalities

We shall start with Markov's inequality as it is the more general of the two results, although historically Chebyshev's inequality came first.[35] Markov's inequality states:

> If Y is any non-negative random variable such that $\mathrm{E}[Y] < \infty$ then, for any real $\epsilon > 0$,
>
> $$\Pr(Y \geq \epsilon) \leq \mathrm{E}[Y]/\epsilon.$$

A simple proof of this result is available as follows:

$$\mathrm{E}[Y] = \mathrm{E}[Y \,|Y < \epsilon] \times \Pr(Y < \epsilon) + \mathrm{E}[Y \,|Y \geq \epsilon] \times \Pr(Y \geq \epsilon).$$

By definition probabilities are greater than or equal to zero and, because $Y \geq 0$, so too must be both conditional expectations. Moreover, $\mathrm{E}[Y \,|Y \geq \epsilon] \geq \epsilon$, because the expectation is a probability weighted average of terms all greater than or equal to ϵ. Hence

$$\mathrm{E}[Y] \geq \mathrm{E}[Y \,|Y \geq \epsilon] \times \Pr(Y \geq \epsilon) \geq \epsilon \times \Pr(Y \geq \epsilon).$$

The final result is then a simple rearrangement of this inequality.

An alternative statement of this result, in the form that we shall use it, is

$$\Pr(g(Z) \geq \epsilon) \leq \frac{\mathrm{E}[g(Z)]}{\epsilon}, \tag{B.1}$$

where $g(Z)$ is a non-negative function of the random variable Z and $\epsilon > 0$.

Chebyshev's inequality is something that people often learn in an introductory statistics course as a way of creating (typically very conservative) confidence intervals. It is implied by Markov's inequality on setting $g(Z) = (Z - c)^2$ for some constant c, giving

$$\Pr\left((Z - c)^2 \geq \epsilon\right) \leq \frac{\mathrm{E}[(Z - c)^2]}{\epsilon} \quad \text{for any real } \epsilon > 0.$$

If we choose $c = \mathrm{E}[Z] \equiv \mu$, so that $\mathrm{E}[(Z - c)^2] = \mathrm{Var}[Z] \equiv \sigma^2$ then we obtain a commonly encountered form of Chebyshev's inequality:[36]

$$\Pr\left((Z - \mu)^2 \geq \epsilon\right) \leq \frac{\sigma^2}{\epsilon}.$$

The event $(Z - \mu)^2 \geq \epsilon$ is that of the Euclidean distance between Z and its mean μ being greater than some arbitrarily small number ϵ. Chebyshev's inequality states that the probability of this event occurring is less than or equal to σ^2/ϵ. The relationship between Chebyshev's inequality and mean square convergence is immediate. If, for some sequence, $\sigma^2 \to 0$ then it follows that the probability of Z being far from its mean goes to zero.

There are other forms of this result that are slightly easier to interpret. For example, there always exists a $k > 0$ such that $\epsilon = k^2\sigma^2$. Making this substitution and noting that $(Z - \mu)^2 \geq k^2\sigma^2$ and $|Z - \mu| \geq k\sigma$ are the same events, we have

$$\Pr(|Z - \mu| \geq k\sigma) \leq \frac{1}{k^2}, \quad \text{for any real } k > 0,$$

[35] Yes, this is the Markov of Gauss-Markov fame. Amongst other things, including being a great mathematician, he was a student of Chebyshev.

[36] Note that Chebyshev's inequality assumes that σ^2 exists, so that $\sigma^2 < \infty$.

or

$$\Pr\left(|Z-\mu| < k\sigma\right) > 1 - \frac{1}{k^2}, \quad \text{for any real } k > 0. \tag{B.2}$$

A simple rearrangement of (B.2) yields the most easily interpreted statement of Chebyshev's inequality:

$$\Pr\left(|Z-\mu| < k\sigma\right) = \Pr\left(\mu - k\sigma < Z < \mu + k\sigma\right) > 1 - \frac{1}{k^2}, \quad \text{for any real } k > 0.$$

In words this last result says that the probability of a random variable falling within k standard deviations of its mean is at least $1 - \frac{1}{k^2}$. For example, there is at least a 75% chance of a random variable lying within two standard deviations of its mean. That such intervals can be very conservative is illustrated by the fact that there is actually a 95% chance that a normally distributed random variable will lie within 1.96 standard deviations of its mean.

B.5 Hölder's Inequality

Hölder's inequality is a generalization of the Cauchy-Schwarz inequality. It states that for $1/p + 1/q = 1$

$$\mathrm{E}\left[|XY|\right] \leq \left(\mathrm{E}\left[|X|^p\right]\right)^{1/p} \left(\mathrm{E}\left[|Y|^q\right]\right)^{1/q}.$$

There are several variations on this theme, including:

$$\sum_{k=1}^{n} |x_k y_k| \leq \left(\sum_{k=1}^{n} |x_k|^p\right)^{1/p} \left(\sum_{k=1}^{n} |y_k|^q\right)^{1/q}$$

$$\sum_{n=1}^{\infty} |x_n \cdot y_n| \leq \left(\sum_{n=1}^{\infty} |x_n|^p\right)^{1/p} \left(\sum_{n=1}^{\infty} |y_n|^q\right)^{1/q}$$

$$\left|\int f(x)g(x)\,dx\right| \leq \left(\int |f(x)|^p\,dx\right)^{1/p} \left(\int |g(x)|^q\,dx\right)^{1/q}.$$

The Cauchy-Schwarz inequality comes about for the special case of $p = q = 2$. See also the Covariance inequality (Section B.3).

B.6 Jensen's Inequality

Jensen's inequality states that if $g(X_n)$ is a concave function of X_n, then

$$\mathrm{E}\left[g(X_n)\right] \leq g\left(\mathrm{E}\left[X_n\right]\right),$$

with strict inequality obtaining if the function g is strictly concave. For example, $\mathrm{E}\left[\ln(X_n)\right] < \ln\left(\mathrm{E}\left[X_n\right]\right)$.

Conversely, if $g(X_n)$ is a convex function of X_n, then

$$\mathrm{E}\left[g(X_n)\right] \geq g\left(\mathrm{E}\left[X_n\right]\right).$$

Again, the strict inequality applies if g is strictly convex.

Finally, there is a corresponding result involving sums which states that, for any real convex function g and positive weights p_i,

$$g\left(\frac{\sum_{i=1}^n p_i x_i}{\sum_{i=1}^n p_i}\right) \leq \frac{\sum_{i=1}^n p_i g(x_i)}{\sum_{i=1}^n p_i};$$

and the inequality is reversed if g is concave.

As a particular case, if the weights p_i are all equal to unity, then

$$g\left(\frac{\sum_{i=1}^n x_i}{n}\right) \leq \frac{\sum_{i=1}^n g(x_i)}{n}.$$

For instance, the $\log(x)$ function is concave so, substituting $g(x) = -\log(x)$ in the previous formula, this establishes (the logarithm of) the familiar arithmetic mean-geometric mean inequality:

$$\frac{x_1 + x_2 + \cdots + x_n}{n} \geq \sqrt[n]{x_1 x_2 \cdots x_n}.$$

As a further example, let us define the weighted harmonic mean to be

$$x_H = \frac{\sum_{i=1}^n w_i}{\sum_{i=1}^n \frac{w_i}{x_i}} = \frac{1}{\sum_{i=1}^n \frac{p_i}{x_i}}, \quad \text{where} \quad 0 \leq p_i = \frac{w_i}{\sum_{i=1}^n w_i} \leq 1 \quad \text{and} \quad \sum_{i=1}^n p_i = 1,$$

with the unweighted harmonic mean, or simply the harmonic mean, to be that special case where the weights are equal, so that $p_i = n^{-1}$ for all $i = 1, \ldots, n$. Then, by Jensen's inequality,

$$\log\left(\frac{1}{x_H}\right) = \log\left(\sum_{i=1}^n \frac{w_i}{x_i}\right) \geq \sum_{i=1}^n w_i \log\left(\frac{1}{x_i}\right) = -\sum_{i=1}^n w_i \log x_i.$$

Since $-\sum_{i=1}^n w_i \log x_i = \log\left(x_1^{-w_1} \ldots x_n^{-w_n}\right)$, it follows that in the special case of the unweighted harmonic mean $\log\left(\frac{1}{x_H}\right) \geq \log\left(\sqrt[n]{x_1^{-1} \ldots x_n^{-1}}\right) = \log\left(\frac{1}{x_G}\right)$, or that $x_G \geq x_H$, where x_G denotes the geometric mean.

B.7 Minkowski's Inequality

Minkowski's inequality is a generalization of the triangle inequality. Like many of these results it has appeared in numerous forms, including

$$\left(\mathrm{E}\left[|X + Y|\right]\right)^{1/p} \leq \left(\mathrm{E}\left[|X|^p\right]\right)^{1/p} + \left(\mathrm{E}\left[|Y|^p\right]\right)^{1/p}, \quad 1 < p < \infty,$$

where equality holds if $X = kY$ for $k > 0$. Alternatively we also have

$$\left(\sum_{k=1}^n |x_k + y_k|^p\right)^{1/p} \leq \left(\sum_{k=1}^n |x_k|^p\right)^{1/p} + \left(\sum_{k=1}^n |y_k|^p\right)^{1/p}.$$