1 Stochastic convergence

The asymptotic theory of minimization estimators relies on various theorems from mathematical statistics. The objective of this section is to explain the main theorems that underpin the asymptotic theory for minimization estimators.

1.1 Infimum and supremum, and the limit operations

1.1.1 Supremum and infimum

The supremum of a set $A$, if it exists, is the smallest number $y$ for which $x \leq y$ for every $x \in A$. Analogously, the infimum of a set $A$, if it exists, is the largest number $y$ for which $x \geq y$ for every $x \in A$. For an open set such as $A = (0, 1)$, the supremum $\sup A$ and the infimum $\inf A$ are well-defined as 1 and 0, respectively. If the supremum is attained for some element in $A$, the supremum is also called the maximum; analogously, the minimum of a set is defined as the infimum if it is attained for some element of $A$. Therefore, the maximum and minimum of $A = (0, 1)$ are not defined, because they are attained at 0 and 1, which are not inside $(0, 1)$.

If we define the extended real line as $\mathbb{R} \cup \{-\infty, +\infty\}$, the sup and inf are always defined on the extended real line.

*Department of Economics, Ohio State University, 429 Arps Hall, Columbus, OH 43210, email dejong@econ.ohio-state.edu
1.1.2 The limit operation

For a deterministic sequence $x_n$, we say that it “converges to $x$”, notation $\lim_{n \to \infty} x_n = x$, if for all $\varepsilon > 0$, there exists an $N_\varepsilon$ such that $|x_n - x| < \varepsilon$ for all $n \geq N_\varepsilon$. The “limsup” (superior limit, “limes superior” in Latin) is defined as

$$\limsup_{n \to \infty} x_n = \lim_{n \to \infty} \sup_{m \geq n} x_m.$$ 

The limsup is the “eventual upper limit” of a sequence. For example, consider $x_n = n^{-1}$. Then

$$\limsup_{n \to \infty} x_n = \sup_{n \geq 1} \inf_{m \geq n} x_m = \inf_{n \geq 1} n^{-1} = 0,$$

just as one would expect. From the earlier discussion, it is now clear that the limsup is always defined on the extended real line. The limsup of a sequence is the “eventual upper bound” of a sequence. For example, $\limsup_{n \to \infty} \sin(n) = 1$, while the limit is obviously undefined. Analogously, we define

$$\liminf_{n \to \infty} x_n = \lim_{n \to \infty} \inf_{m \geq n} x_m.$$ 

For $x_n = n^{-1}$, we obtain

$$\liminf_{n \to \infty} x_n = \inf_{n \geq 1} \sup_{m \geq n} x_m = \sup_{n \geq 1} 0 = 0,$$

as one would expect from an “eventual lower limit”. If and only if the limsup and liminf of a sequence are equal, then the limit of the sequence exists.

If, after noting that $\log(1 + x) \leq x$, we read something like

$$\lim_{n \to \infty} n^2 \log(\Phi(n)) \leq \lim_{n \to \infty} n^2(\Phi(n) - 1) = 0$$

where $\Phi(\cdot)$ denotes the standard normal distribution function, we should realize the following. In principle, when writing down the left-hand side expression, it is possible for this expression to be undefined. The correct interpretation of the above equation is that the inequality finds an upper bound if the expression exists; but formally the existence of the left-hand side remains unproven. The assertion that the lim sup is less than or equal to zero is always true however if the limsup is defined on the extended real line.
1.2 Almost sure convergence, convergence in probability, and convergence in distribution

1.2.1 Definitions of stochastic convergence concepts

While for a deterministic sequence $X_n$ it is relatively straightforward to define convergence of $X_n$ to a number $X$, the situation becomes more complicated in the case where $X_n$ and/or $X$ is allowed to be a random variable. There are four types of convergence for random variables that are used frequently in the econometric literature. They are convergence in distribution, convergence in probability, convergence in $L_p$, and almost sure convergence. We say that $X_n$ converges to $X$ in distribution if

$$\lim_{n \to \infty} P(X_n \leq y) = P(X \leq y)$$

for all $y$ for which $P(X \leq y)$ is continuous in $y$. To see why we only want this requirement to hold only for continuity points of $P(X \leq y)$, consider $X_n = n^{-1}$. This is a deterministic sequence and we should hope that it converges in distribution to $X = 0$. However, setting $y = 0$,

$$P(X_n \leq 0) = P(n^{-1} \leq 0) = 0 \neq P(X \leq 0) = 1.$$ 

Because the convergence in distribution requirement excludes discontinuity points however, $X_n = n^{-1}$ does converge in distribution to zero.

$X_n$ converges in probability to $X$ if for all $\delta > 0$,

$$\lim_{n \to \infty} P(|X_n - X| > \delta) = 0.$$ 

We typically write “$X_n \xrightarrow{p} X$” or “$\text{plim}_{n \to \infty} X_n = X$” to denote convergence in probability. $X_n$ is said to converge almost surely to $X$ if

$$P(\lim_{n \to \infty} X_n = X) = 1$$

which is equivalent to the condition that for all $\delta > 0$,

$$\lim_{n \to \infty} P(\sup_{m \geq n} |X_m - X| > \delta) = 0.$$ 

We then write “$X_n \xrightarrow{\text{as}} X$” or “$\lim_{n \to \infty} X_n = X$ almost surely”. Convergence in $L_p$ of $X_n$ to $X$ for $p \geq 1$ holds if

$$\lim_{n \to \infty} E|X_n - X|^p = 0.$$ 

3
1.2.2 Differences between stochastic convergence concepts

There are subtle differences among these convergence concepts; however, convergence in distribution is in some ways in a different category from the other stochastic convergence concepts. The important thing to realize about convergence in distribution is that if \( X_n \) converges to \( X \), and \( X \) is a standard normally distributed random variable, then \( X_n \) also converges to \( Y \), which is a random variable that is also standard normally distributed. For example, a sequence \( X_n \) of i.i.d. \( N(0, 1) \) random variables converges in distribution to a \( N(0, 1) \) random variable as \( n \to \infty \). This may be counterintuitive, because \( X_n \) will never get “closer” to any particular random variable; but the important thing to realize is that only the distribution of \( X_n \) needs to converge for convergence in distribution.

While convergence in distribution only says something about distributions, convergence in probability, almost sure convergence and convergence in \( L_p \) are statements implying that \( X_n \) and \( X \) are asymptotically “close”. Almost sure convergence implies convergence in probability, as can be easily seen from my definitions above. Convergence in \( L_p \) implies convergence in probability. This follows from the Markov inequality \( P(Y > \delta) \leq \delta^{-1}E|Y| \), because

\[
P(|X_n - X| > \delta) = P(|X_n - X|^p > \delta^p) \leq \delta^{-p}E|X_n - X|^p.
\]

Convergence in probability implies convergence in distribution, and convergence in distribution to a constant implies convergence in probability to that constant; I will leave this unproven here. No other relations between concepts hold without further assumptions.

1.2.3 An odd example

A classical example of a sequence \( X_n \) that converges in probability to zero but not in \( L_p \) is the following. Let \( X_n \) be a random variable that takes the value \( \exp(n) \) with probability \( 1/n \) and zero otherwise. Then for all \( \delta > 0 \),

\[
\lim_{n \to \infty} P(|X_n - X| > \delta) \leq \lim_{n \to \infty} P(X_n = \exp(n)) = \lim_{n \to \infty} n^{-1} = 0,
\]

so \( X_n \) converges in probability to 0. However, for any \( p \geq 1 \),

\[
E|X_n|^p = \exp(np)P(X_n = \exp(n)) = n^{-1}\exp(np),
\]

so \( X_n \) does not converge in \( L_p \). Another example where convergence in probability holds but convergence in \( L_p \) does not is the case where \( X_n = Y/n \) where \( Y \) is Cauchy. In that case, we have convergence in probability, but not in \( L_p \), because moments of order 1 and higher of the Cauchy distribution do not exist.
If in addition $X_n$ is independent across $n$, we also have an example of a sequence that converges in probability, but not almost surely. This is because while $X_n \xrightarrow{p} 0$, $\sup_{m \geq n} X_m$ does not converge to zero in probability. After all, $\sup_{m \geq n} X_m \geq \exp(n)$ if one of the $X_m$ for which $m \geq n$ is nonzero; that is, $\sup_{m \geq n} X_m \leq 1$ only if $X_m = 0 \forall m \geq n$. Therefore, for all $0 < \delta < 1$,

$$P(\sup_{m \geq n} |X_m| \leq \delta) \leq P(\forall m \geq n : X_m = 0)$$

$$= \prod_{m=n}^{\infty} P(X_m = 0) = \prod_{m=n}^{\infty} (1 - 1/m)$$

$$= \exp(\sum_{m=n}^{\infty} \log(1 - 1/m)) = 0.$$

To show that last result, note that because $\log(1 + x) \leq x$ for $x > -1$,

$$\lim_{K \to \infty} \exp(\sum_{m=n}^{K} \log(1 - 1/m)) \leq \lim_{K \to \infty} \exp(- \sum_{m=n}^{K} (1/m)) = 0$$

because $m^{-1}$ is not summable from 1 to infinity.

This implies that $X_n$ converges in probability to zero, but not almost surely.

1.2.4 Averages

Less exotic statistics that we will encounter in econometrics are averages of i.i.d. or weakly dependent random variables. Consider the following simple example. Let

$$X_n = n^{-1/2} \sum_{i=1}^{n} (z_i - E z_i),$$

where the $z_i$ are i.i.d. and $E z_i^2 < \infty$. Then $X_n$ converges in distribution to a normally distributed random variable. For each deterministic sequence $a_n$ such that $\lim_{n \to \infty} a_n = 0$ we have $a_n X_n \xrightarrow{p} 0$. However, it turns out that $a_n X_n \xrightarrow{a.s.} 0$ is only true if $a_n (\log \log n)^{1/2} \to 0$; this follows from the so-called law of the iterated logarithm.
1.3 The law of large numbers

1.3.1 Statement of the law of large numbers

Consistency of estimators is derived using a law of large numbers (LLN). The LLN asserts that

\[ n^{-1} \sum_{i=1}^{n} (z_i - E z_i) \]

converges to zero. It can be shown that if \( E |z_i| < \infty \) and \( z_i \) is i.i.d. (independent and identically distributed), then the above expression converges to zero almost surely (and therefore also in probability; the almost sure version is Kolmogorov’s strong law of large numbers). If \( E |z_i| \) is not finite, the LLN does not need to hold; a counterexample would be the Cauchy distribution, for which \( \bar{z}_n \) is distributed identically to \( z_i \). If convergence to zero holds almost surely, we talk about the strong law of large numbers, while if the converge to zero is in probability, we refer to the result as a weak law of large numbers.

1.3.2 Weak LLN under independence

If \( z_i \) has mean zero, is i.i.d. and \( E z_i^2 < \infty \), then

\[ V(n^{-1} \sum_{i=1}^{n} z_i) = n^{-2} \sum_{i=1}^{n} V(z_i) = n^{-1} V(z_i) \to 0, \]

so a weak law of large numbers will hold for \( z_i \) in this case. The condition \( E z_i^2 < \infty \) can be weakened to \( E |z_i| < \infty \) by a truncation argument. This can be done as follows. Define \( z_i^K = z_i I(|z_i| > K) \) and \( z_iK = z_i I(|z_i| \leq K) \). Then

\[
E|n^{-1} \sum_{i=1}^{n} z_i| \leq E|n^{-1} \sum_{i=1}^{n} (z_iK - E z_iK)| + E|n^{-1} \sum_{i=1}^{n} (z_i^K - E z_i^K)|
\]

\[
\leq (E|n^{-1} \sum_{i=1}^{n} (z_iK - E z_iK)|^2)^{1/2} + 2n^{-1} \sum_{i=1}^{n} E|z_i^K|
\]

\[
\leq (n^{-1} V(z_iK))^1/2 + 2E|z_i|I(|z_i| > K)
\]

\[
\leq (n^{-1} K^2)^{1/2} + 2 \int_{|z|>K} |z| dF(z),
\]

and by taking first the limit as \( n \to \infty \) and then the limit as \( K \to \infty \), the convergence in \( L_1 \) follows.

To prove Kolmogorov’s strong law of large numbers is much more involved; this typically requires a result known as the three series theorem. I will not pursue that here.
1.3.3 Discussion

There exist many formulations of the LLN that are different from the above. For example, if $z_i$ is uncorrelated, $Ez_i^2 < \infty$ and

$$\text{Var}(n^{-1} \sum_{i=1}^{n} (z_i - Ez_i)) = n^{-2} \sum_{i=1}^{n} \text{Var}(z_i) \to 0,$$

then the weak LLN will also hold, in spite of the fact that the i.i.d. assumption has not been made. Also, the independence assumption can be relaxed further to allow for random variables $z_i$ that have “fading memory” properties. To illustrate this, note that in general

$$\text{Var}(n^{-1} \sum_{i=1}^{n} (z_i - Ez_i)) = n^{-2} \sum_{i=1}^{n} \sum_{s=1}^{n} \text{cov}(z_i, z_s)$$

$$\leq 2n^{-2} \sum_{i=1}^{n} \sum_{s=1}^{n} |\text{cov}(z_i, z_s)|$$

$$\leq 2n^{-2} \sum_{i=1}^{n} \sum_{m=0}^{\infty} |\text{cov}(z_i, z_{i+m})|,$$

and if the last expression converges to zero somehow, then a weak LLN can be obtained as well.

Often, alternatives to the classical LLN for i.i.d. variables involve assumptions on the existence of moments $E|z_i|^{1+\delta}$ for some $\delta > 0$.

1.4 The central limit theorem

1.4.1 Statement of the central limit theorem

The central limit theorem (CLT) states that

$$n^{-1/2} \sum_{i=1}^{n} (z_i - Ez_i) \xrightarrow{d} N(0, \text{Var}(z_i)).$$

If $z_i$ is i.i.d. and $Ez_i^2 < \infty$, then the above result holds.
1.4.2 CLT and characteristic function

The characteristic function of a random variable $X$ is defined as

$$\psi(r) = E \exp(irX),$$

where $i$ satisfies $i^2 = -1$. There is a one-to-one correspondence between characteristic functions and distribution functions, and because $|\psi(\cdot)| \leq 1$, showing that the characteristic function converges to the characteristic function of a normal distribution suffices to show convergence in distribution to a normal. The reasoning for proving a CLT for a mean zero, i.i.d. sequence $z_i$ with a variance $\sigma^2$ then starts by observing that

$$E \exp(irn^{-1/2} \sum_{i=1}^{n} z_i) = \prod_{i=1}^{n} E \exp(irn^{-1/2} z_i)$$

by independence. By the Taylor series expansion, $\exp(x) \approx 1 + x + 0.5x^2$ and $\log(1 + x) \approx x$, and therefore

$$E \exp(irn^{-1/2} \sum_{i=1}^{n} z_i) \approx \prod_{i=1}^{n} E(1 + (irn^{-1/2} z_i) + 0.5(irn^{-1/2} z_i)^2)$$

$$= \prod_{i=1}^{n} (1 - 0.5r^2 n^{-1}\sigma^2) = \exp(\sum_{i=1}^{n} \log(1 - 0.5r^2 n^{-1}\sigma^2))$$

$$\approx \exp(-0.5r^2\sigma^2),$$

which is the characteristic function of the $N(0, \sigma^2)$ distribution. Of course, the analytical difficulty is to justify the “$\approx$” signs in the above reasoning.

1.4.3 Discussion

For the CLT, lots of alternative specifications are also possible. It is possible to allow for some heterogeneity of the distributions of $z_i$, and also the requirement of independence can be relaxed to some type of “fading memory” assumption. “Heterogeneity of the distributions” of $z_i$ refers to the fact that the distributions $F_i(x)$ of $z_i$ are not identical for all $i$. 
1.5 Moment conditions

Many articles and books in econometrics contain so-called moment conditions of the type

\[ E|z_i| < \infty, \quad \text{or} \quad \limsup_{n \to \infty} n^{-1} \sum_{i=1}^{n} E|z_i|^2 < \infty, \]

or assume that for some \( \delta > 0 \),

\[ E|z_i|^{2+\delta} < \infty. \]

By now, you probably realize that these conditions typically originate from the verification of conditions necessary for obtaining a law of large numbers or central limit theorem. For example, \( n^{-1} \sum_{i=1}^{n} (z_i - E z_i) \overset{a.s.}{\to} 0 \) if the \( z_i \) are i.i.d. and \( E|z_i| < \infty \). So, if somewhere along the line we need the result \( n^{-1} \sum_{i=1}^{n} (z_i - E z_i) \overset{a.s.}{\to} 0 \), we simply impose \( E|z_i| < \infty \) as a regularity condition. The same holds for the central limit theorem. Finally, note that \( E|z_i|^p < \infty \) is implied by \( E|z_i|^q < \infty \) for some \( q > p \); this is because for \( 1 \leq p \leq q \),

\[ (E|X|^p)^{1/p} \leq (E|X|^q)^{1/q}. \]

An example of a distribution for which moment conditions can fail to hold is the Student distribution. If \( z_i \) has a Student distribution with \( k \) degrees of freedom, then \( E|z_i|^k \) does not exist; remember that a \( t \) distribution with one degree of freedom is the Cauchy distribution.

1.6 The dominated convergence theorem

A tool that is frequently used in asymptotic theory in econometrics is the dominated convergence theorem.

**Theorem 1** If \( X_n \overset{p}{\to} X \) and \(|X_n| \leq Y\) a.s. for some random variable \( Y \) and \( EY < \infty \), then

\[ \lim_{n \to \infty} EX_n = EX. \]

This result is widely used in econometrics; notably, it is used for asymptotic normality of minimization estimators when the mean value in the Taylor series expansion argument is replaced by its limit.
1.7 A joint convergence result

1.7.1 Justification of the “delta method”

The following result plays a key role in asymptotic theory for justifying variance estimation:

**Theorem 2** If $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{p} c$ for some constant $c$ and $f(x, y)$ is continuous in both arguments, then $f(X_n, Y_n) \xrightarrow{d} f(X, c)$.

Using the above theorem, we can reason as follows. If $n^{-1} \sum_{i=1}^n z_i^2 \xrightarrow{p} \sigma^2$ and $n^{-1/2} \sum_{i=1}^n z_i \xrightarrow{d} N(0, \sigma^2)$, then

$$\frac{n^{-1/2} \sum_{i=1}^n z_i}{\sqrt{n^{-1} \sum_{i=1}^n z_i^2}} \xrightarrow{d} N(0, 1).$$

Note that this implies that $f(x, y) = x/\sqrt{y}$, which is not continuous for $y$ near 0, leaving a small problem in our argument. This problem can be formally fixed, using the assumption that $\sigma^2 > 0$.

1.7.2 A beginner’s mistake

However, it is **not** allowed to reason

$$\frac{n^{-1/2} \sum_{i=1}^n z_i}{\sqrt{n^{-1} \sum_{i=1}^n z_i^2}} = n^{1/2} \frac{n^{-1} \sum_{i=1}^n z_i}{\sqrt{n^{-1} \sum_{i=1}^n z_i^2}} \xrightarrow{d} n^{1/2} \frac{0}{\sigma^2} = 0$$

as many beginners tend to do. You cannot find the continuous function $f(., .)$ that will allow you to do this trick; you would need $f(x, y) = n^{1/2} x/\sqrt{y}$, but dependence of $f(., .)$ on $n$ is not allowed. The correct reasoning displayed previously assumed $f(x, y) = x/\sqrt{y}$ and clearly, this is allowed.

1.8 The multivariate case

Often when convergence in distribution of a random vector is shown, a result know as the Cramèr-Wold device is used. The Cramèr-Wold device is a tool that can be used to simplify the problem of showing convergence in distribution of a vector to convergence in distribution of scalars. It asserts that for a vector $X_n$ of dimension $k$, $X_n \xrightarrow{d} X$ if and only if for all $k$-vectors $\xi$, $\xi'X_n \xrightarrow{d} \xi'X$. This implies that using the Cramèr-Wold device, the problem of showing asymptotic normality of a vector can be translated into a problem involving an application of the CLT for scalars.
1.9 Why asymptotic normality?

The reason econometricians are interested in establishing asymptotic normality of estimators is that this property allows us to justify inference procedures such as $t$-tests and $F$-tests. Of course, $t$-tests that are justified based on asymptotic theory are really asymptotically standard normal, and $F$-tests that are justified based on asymptotic theory are really asymptotically chi-squared. Asymptotic normality of a $k$-vector $X_n$, i.e. $X_n \xrightarrow{d} N(0, V)$, implies that if $\hat{V} \xrightarrow{p} V$,

$$\frac{X_{nj}}{\sqrt{\hat{V}_{jj}}} \xrightarrow{d} N(0, 1),$$

which will imply validity of $t$-values. A similar argument can be used to show that $F$-tests are asymptotically valid if we have asymptotic normality of our estimator.

1.10 The “argmin” functional

The “argmin” is defined as a value at which the function over which the argmin is taken is minimized. For example, $\text{argmin}_{x \in \mathbb{R}}(1 + x^2) = 0$. Note that for two functions that are arbitrarily close, the argmin functional can be very different. This will be a source of complications when we will prove consistency of estimators in situations when we cannot explicitly write $\hat{\theta}_n$ as an explicit and relatively simple function of the $(y_i, x_i), \ i = 1, \ldots, n$. Consider for example, for $\theta \in [0, 1]$,

$$f_n(\theta) = \theta n^{-1}$$

and

$$g_n = (1 - \theta)n^{-1}.$$

Then $\sup_{\theta \in \Theta} |f_n(\theta) - g_n(\theta)| \to 0$, while $\text{argmin}_{\theta \in [0, 1]} f_n(\theta) = 0$ and $\text{argmin}_{\theta \in [0, 1]} g_n(\theta) = 1$. Obviously, both argmins do not get close asymptotically; however, also note that this is because the limit function 0 that both $f_n(\cdot)$ and $g_n(\cdot)$ converge to does not have a unique minimum.

The “argmax” is defined analogously to the “argmin”, as the value where at which the function over which the argmax is taken is maximized. In principle, the argmin of a function $Q_n(\theta)$ can be attained at multiple values in $\Theta$, which would make the statement $\hat{\theta}_n = \text{argmin}_{\theta \in \Theta} Q_n(\theta)$ undefined; but this possibility is typically disregarded in econometric prose.
1.11 Exercises

1. Assume that \( x_n \) is a nondecreasing, bounded deterministic sequence of scalars. Show that \( \lim_{n \to \infty} x_n \) exists.

2. Assume that \( X_n \) is an increasing sequence of random variables such that \( X_n \xrightarrow{p} X \). Show that in this case, \( X_n \xrightarrow{p} X \) also.

3. Assume that \( X_n \) and \( Y_n \) are sequences of random variables such that \( X_n \xrightarrow{p} X \) and \( Y_n \xrightarrow{p} Y \). Show that \( Y_n + X_n \xrightarrow{p} X + Y \).

4. Assume that \( X_n \) and \( Y_n \) are sequences of random variables such that \( X_n \xrightarrow{d} X \) and \( Y_n \xrightarrow{d} Y \). Find an example that shows that \( Y_n + X_n \xrightarrow{d} X + Y \) does not need to hold.

5. Assume that \( X_n \) is a sequence of random variables such that \( X_n \xrightarrow{d} c \) for a constant \( c \). Show that \( X_n \xrightarrow{p} c \) also.

6. In section 1.2.3, what happens to the “odd example” if the \( n^{-1} \) in its definition is replaced by \( n^{-2} \)?

7. Let \( x_i \) be i.i.d. and \( N(0,1) \) distributed. Using characteristic functions, show that \( n^{-1/2} \sum_{i=1}^{n} x_i \) is also \( N(0,1) \) distributed.

8. Formally prove the dominated convergence theorem.
2 Asymptotics of the linear model: the scalar case

2.1 The consistency argument

In this section we will investigate the asymptotic behavior of least squares estimators. The purpose is only to serve as an introduction to the more general case, where we will only assume that our estimator is some minimization estimator, for which no closed form solution is available. To understand the principles involved better, we will focus on the case of a scalar regressor $x_i$ in this section.

In the case of the simple linear model

$$y_i = \theta_0 x_i + \varepsilon_i,$$

where $x_i \in \mathbb{R}$, the closed form solution for the least squares estimator is

$$\hat{\theta}_n = \arg\min_{\theta \in \Theta} n^{-1} \sum_{i=1}^{n} (y_i - \theta x_i)^2$$

$$= \frac{\sum_{i=1}^{n} y_i x_i}{\sum_{i=1}^{n} x_i^2}$$

if the parameter space $\Theta$ is equal to $\mathbb{R}$. Parameter spaces other than $\mathbb{R}$ can sometimes be natural; for example, in an ARCH(1) model, a parameter space of $[0, 1]$ for the $\alpha$ parameter may be preferable. Later, we will see that the general consistency proof forces an assumption of compactness of the parameter space on us.

Note that this solution for $\hat{\theta}_n$ was obtained by differentiation with respect to $\theta$. Next, we note that - if the true model is correct - it is possible to write

$$\hat{\theta}_n = \theta_0 + \frac{\sum_{i=1}^{n} \varepsilon_i x_i}{\sum_{i=1}^{n} x_i^2}.$$

After this step, we have to make some assumption on the behavior of $x_i$: is it a deterministic or a stochastic sequence?

2.1.1 Consistency argument: the deterministic regressor case

First, we assume that $x_i$ is deterministic. Then note that

$$\frac{\sum_{i=1}^{n} \varepsilon_i x_i}{\sum_{i=1}^{n} x_i^2} = \frac{n^{-1} \sum_{i=1}^{n} \varepsilon_i x_i}{n^{-1} \sum_{i=1}^{n} x_i^2},$$

and therefore $\hat{\theta}_n \xrightarrow{as} \theta_0$ if
1. $n^{-1} \sum_{i=1}^{n} \varepsilon_i x_i \xrightarrow{as} 0$;
2. $n^{-1} \sum_{i=1}^{n} x_i^2 \xrightarrow{p} Q$, where $Q > 0$.

2.1.2 Consistency argument: the random regressor case

Under the assumption that $x_i$ is random, sufficient conditions for $\hat{\theta}_n \xrightarrow{as} \theta_0$ are

1. $n^{-1} \sum_{i=1}^{n} \varepsilon_i x_i \xrightarrow{as} 0$;
2. $n^{-1} \sum_{i=1}^{n} x_i^2 \xrightarrow{as} Q$, where $Q > 0$.

Both conditions follow, by Kolmogorov’s law of large numbers, from the conditions

1. $(x_i, \varepsilon_i)$ is i.i.d.;
2. $Ex_i^2 < \infty$;
3. $E\varepsilon_i x_i = 0$;
4. $E|\varepsilon_i x_i| < \infty$.

In the case that $\varepsilon_i$ and $x_i$ are independent of each other, conditions 1-4 are implied by

1. $(x_i, \varepsilon_i)$ is i.i.d.;
2. $Ex_i^2 < \infty$;
3. $E\varepsilon_i = 0$;
4. $E|\varepsilon_i| < \infty$.

2.2 The asymptotic normality argument

Finding the limit distribution of $\hat{\theta}_n$ requires some more work than the simple consistency result. If we assume that the $\varepsilon_i$ are i.i.d. normal with variance $\sigma^2$ and $x_i$ is deterministic, then

$$\frac{\sum_{i=1}^{n} \varepsilon_i x_i}{\sum_{i=1}^{n} x_i^2}$$

is distributed $N(0, \sigma^2(\sum_{i=1}^{n} x_i^2)^{-1})$ for any sample size. Therefore, if $n^{-1} \sum_{i=1}^{n} x_i^2 \xrightarrow{p} Q$, we get

$$n^{1/2}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(0, \sigma^2 Q^{-1})$$.

The asymptotic version of this argument simply replaces the exact normality by asymptotic normality, and uses a CLT as a replacement for the exact normality assumption.
2.2.1 Asymptotic normality argument: the deterministic regressor case

Sufficient conditions for the asymptotic normality result \( n^{1/2}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(0, \sigma^2 Q^{-1}) \) are

1. \( n^{-1/2} \sum_{i=1}^{n} \varepsilon_i x_i \xrightarrow{d} N(0, \sigma^2 Q) \);
2. \( n^{-1} \sum_{i=1}^{n} x_i^2 \xrightarrow{p} Q \) where \( Q > 0 \).

It is possible to use a central limit theorem for independent, but not identically distributed random variables to establish 1.

2.2.2 Asymptotic normality argument: the random regressor case

At this point, our interest will be in relaxing the assumption that the \( x_i \) are deterministic. If we assume

1. \( (x_i, \varepsilon_i) \) is i.i.d.;
2. \( E\varepsilon_i x_i = 0; \)
3. \( E\varepsilon_i^2 x_i^2 < \infty \) and \( E\varepsilon_i^2 x_i^2 < \infty; \)

then by the central limit theorem,

\[
n^{-1/2} \sum_{i=1}^{n} x_i \varepsilon_i \xrightarrow{d} N(0, E\varepsilon_i^2 x_i^2),
\]

and in addition,

\[
n^{-1} \sum_{i=1}^{n} x_i^2 \xrightarrow{p} Q,
\]

and therefore by Theorem 2,

\[
\frac{n^{-1/2} \sum_{i=1}^{n} x_i \varepsilon_i}{n^{-1} \sum_{i=1}^{n} x_i^2} \xrightarrow{d} N(0, Q^{-2} E\varepsilon_i^2 x_i^2).
\]

Therefore,

\[
n^{1/2}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(0, Q^{-2} P),
\]

where \( P = E\varepsilon_i^2 x_i^2. \)
2.2.3 Heteroskedasticity

If we assume in addition that $\varepsilon_i$ and $x_i$ are independent, then

$$P = E\varepsilon_i^2 x_i^2 = E\varepsilon_i^2 E x_i^2 = \sigma^2 Q,$$

and then our result becomes

$$n^{1/2}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(0, \sigma^2 Q^{-1}).$$

The above result also follows from assuming that $E(\varepsilon_i^2 | x_i) = \sigma^2$ (a homoskedasticity assumption). However, please note that asymptotic normality is also obtained without homoskedasticity assumption. We will discuss this point further when we discuss the multivariate case.

In the sections that follow, we will always assume that the $x_i$ sequence is stochastic.
3 Misspecified linear models: the scalar case

Unfortunately, in practice, we never know whether the model that we specified is true or not. Therefore, we may wonder what happens to our estimator if the model estimated is misspecified. For example, if

\[ y_i = x_i^2 + \varepsilon_i \]

where \((\varepsilon_i, x_i)\) is i.i.d. and \(\varepsilon_i\) is independent of \(x_i\), is the true model instead of the linear model from the previous section, what will our OLS estimator do? Will it wander off to infinity, or will it have a limit distribution?

3.1 Misspecified linear models: consistency

In order to shed some light on this issue, note that for the OLS estimator from the previous section we know that

\[
\hat{\theta}_n = \frac{\sum_{i=1}^{n} y_i x_i}{\sum_{i=1}^{n} x_i^2}
\]

\[
= \frac{\sum_{i=1}^{n} x_i^3}{\sum_{i=1}^{n} x_i^2} + \frac{\sum_{i=1}^{n} \varepsilon_i x_i}{\sum_{i=1}^{n} x_i^2}
\]

\[
= \frac{n^{-1} \sum_{i=1}^{n} x_i^3}{n^{-1} \sum_{i=1}^{n} x_i^2} + \frac{n^{-1} \sum_{i=1}^{n} \varepsilon_i x_i}{n^{-1} \sum_{i=1}^{n} x_i^2}.
\]

By the law of large numbers, if \(E|x_i|^3 < \infty\) and \(E|\varepsilon_i x_i| < \infty\),

\[
n^{-1} \sum_{i=1}^{n} x_i^3 \xrightarrow{as} Ex_i^3,
\]

\[
n^{-1} \sum_{i=1}^{n} \varepsilon_i x_i \xrightarrow{as} 0,
\]

and

\[
n^{-1} \sum_{i=1}^{n} x_i^2 \xrightarrow{as} Q = Ex_i^2.
\]

Therefore,

\[
\hat{\theta}_n \xrightarrow{as} Ex_i^3/Ex_i^2.
\]

The value \(\theta^* = Ex_i^3/Ex_i^2\) is often referred to as the pseudo-true value.
3.2 Misspecified linear models: asymptotic normality

To see what happens to the limit distribution of our misspecified estimator, define $\hat{Q}_n = n^{-1} \sum_{i=1}^{n} x_i^2$ and note that

$$n^{1/2}(\hat{\theta}_n - Ex_i^3/Q) = \frac{n^{-1/2} \sum_{i=1}^{n} \varepsilon_i x_i} {n^{-1} \sum_{i=1}^{n} x_i^2} + \frac{n^{-1/2} \sum_{i=1}^{n} x_i^3} {n^{-1} \sum_{i=1}^{n} x_i^2} - \frac{n^{-1/2} \sum_{i=1}^{n} Ex_i^3} {Q}$$

$$= \frac{Qn^{-1/2} \sum_{i=1}^{n} \varepsilon_i x_i} {Q\hat{Q}_n} + \frac{Qn^{-1/2} \sum_{i=1}^{n} (x_i^3 - Ex_i^3)} {Q\hat{Q}_n} - \frac{n^{1/2}(\hat{Q}_n - Q)Ex_i^3} {Q\hat{Q}_n Q}.$$

Therefore, by Theorem 2, $n^{1/2}(\hat{\theta}_n - Ex_i^3/Q)$ has the same limit distribution as

$$Qn^{-1/2} \sum_{i=1}^{n} \varepsilon_i x_i + Qn^{-1/2} \sum_{i=1}^{n} (x_i^3 - Ex_i^3) - n^{1/2}(\hat{Q}_n - Q)Ex_i^3.$$

In general,

$$(n^{-1/2} \sum_{i=1}^{n} \varepsilon_i x_i, n^{-1/2} \sum_{i=1}^{n} (x_i^3 - Ex_i^3), n^{-1/2} \sum_{i=1}^{n} (x_i^2 - Ex_i^2))$$

is an asymptotically normal vector; therefore, $n^{1/2}(\hat{\theta}_n - Ex_i^3/Q)$ is asymptotically normally distributed.

3.3 Discussion

The striking thing about this result is that $\hat{\theta}_n$ is still both converging at rate $n^{1/2}$ and asymptotically normally distributed, in spite of the obvious misspecification of the model that was estimated. If we define $\theta^* = Ex_i^3/Q$, then our result can be rewritten as

$$n^{1/2}(\hat{\theta}_n - \theta^*) \overset{d}{\to} N(0, V).$$

The value for $\theta^*$ could also be obtained along a different route. Since the OLS estimator is the least squares estimator, one might expect that $\theta^*$ minimizes $E(y_i - \theta x_i)^2$. This is indeed the case, since

$$E(y_i - \theta x_i)^2 = E(\varepsilon_i + x_i^2 - \theta x_i)^2$$

$$= \sigma^2 + Ex_i^4 - 2\theta Ex_i^3 + \theta^2 Ex_i^2 - 2E\varepsilon_i x_i^2 - 2\theta E\varepsilon_i x_i;$$

Differentiation with respect to $\theta$ then learns that $\theta^*$ is the minimizer of $E(y_i - \theta x_i)^2$ if $E\varepsilon_i x_i = 0$. Our conclusion therefore is that if the population objective function is uniquely minimized at some value, we will get consistency for that value. This will turn out to be true in more general settings as well.
4 Asymptotics of the linear model: multivariate case

Now let us consider the well-known model

\[ y_i = \theta_0' x_i + \varepsilon_i \]

where \( \theta, x_i \in \mathbb{R}^k \). The OLS estimator now is

\[ \hat{\theta}_n = (X'X)^{-1} X'y \]

where \( X' = (x_1, x_2, \ldots, x_n) \) and \( y' = (y_1, \ldots, y_n) \) and \( \varepsilon' = (\varepsilon_1, \ldots, \varepsilon_n) \). If the model is true,

\[ \hat{\theta}_n = (X'X)^{-1} X'y = \theta_0 + (X'X)^{-1} X' \varepsilon. \]

4.1 Consistency of the linear model: multivariate case

Since all elements of \( X'X \) are summations over \( n \) elements, it is not unreasonable to assume that \( X'X/n \xrightarrow{as} Q \), where \( Q \) denotes some \( k \times k \) matrix that is assumed to be invertible, and that \( X' \varepsilon \xrightarrow{as} 0 \). These results can be derived by applying the LLN \( k^2 \) and \( k \) times, respectively.

To see why \( X'X \) and \( X'y \) are summations over \( n \) elements, the following trick is useful. Let \( s_i \) denote a vector of all zero elements, except for an element of 1 on spot \( i \). Then

\[ I_n = \sum_{i=1}^n s_is_i' \]

and similarly,

\[ y = \sum_{i=1}^n x_is_i'y. \]

Therefore, because

\[ \hat{\theta} = \theta_0 + (n^{-1} \sum_{i=1}^n x_i x_i')^{-1} n^{-1} \sum_{i=1}^n x_i \varepsilon_i \]

and letting \( |\cdot| \) denote the Euclidean norm, consistency of \( \hat{\theta}_n \) follows if

1. \( (\varepsilon_i, x_i) \) is i.i.d.;
2. \( E\varepsilon_i x_i = 0; \)
3. \( E|x_i|^2 < \infty; \)
4. \( E|\varepsilon_i||x_i| < \infty. \)
4.2 Asymptotic normality of the linear model: multivariate case

Also, by the multivariate central limit theorem, usually

\[ n^{-1/2}X'\varepsilon \xrightarrow{d} N(0, E(X'\varepsilon\varepsilon'X)), \]

and therefore, denoting \( P = E(X'\varepsilon\varepsilon'X/n) = E\varepsilon_i^2x_ix'_i, \)

\[ n^{1/2}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(0, Q^{-1}PQ^{-1}). \]

If in addition we assume that \( E(\varepsilon_i^2|x_i) = \sigma^2 \) (i.e., no heteroskedasticity), then

\[ P = E(X'\varepsilon\varepsilon'X/n) = E(X'E(\varepsilon\varepsilon'|X)X/n) \]
\[ = \sigma^2 E(X'I_nX/n) = \sigma^2Q, \]

where \( I_n \) denotes the \( n \times n \) identity matrix. Then we obtain the more familiar result

\[ n^{1/2}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(0, \sigma^2Q^{-1}). \]

Writing

\[ n^{1/2}(\hat{\theta}_n - \theta_0) = (n^{-1}\sum_{i=1}^{n} x_ix'_i)^{-1}n^{-1/2}\sum_{i=1}^{n} \varepsilon_i x_i, \]

and again letting \( |.| \) denote the Euclidean norm, we can now conclude that we obtain the asymptotic normality result \( n^{1/2}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(0, Q^{-1}PQ^{-1}) \) if

1. \((\varepsilon_i, x_i)\) is i.i.d.;
2. \( E|x_i|^2 < \infty; \)
3. \( E\varepsilon_i x_i = 0; \)
4. \( E\varepsilon_i^2|x_i|^2 < \infty. \)

Note that it is possible to estimate \( Q^{-1}PQ^{-1} \) also. This is done by using \( \hat{Q}_n \) as an estimate for \( Q \) and using

\[ \hat{P}_n = n^{-1}\sum_{t=1}^{n} x_ix'_i\hat{\varepsilon}_i^2, \]

where \( \hat{\varepsilon}_i \) is the regression residual from the regression of \( y_i \) on \( x_i \). If \( t \)-values are calculated in this way, we usually say that “White’s \( t \)-values”, “\( t \)-values according to the method of White”, or “heteroscedasticity-consistent \( t \)-values” are calculated.
5 Consistency of extremum estimators: basic case

The next thing that we will set out to do is the consideration of general minimization estimators for which no closed form solution is available. Estimators are usually proposed as solutions of some minimization problem; maximum likelihood estimators and (non)linear least squares estimators are examples of this. Closed form solutions for $\hat{\theta}_n$ are rarely available, and therefore it is of great interest to obtain general results regarding consistency and asymptotic normality of such estimators. Let

$$\hat{\theta}_n = \arg\min_{\theta \in \Theta} Q_n(\theta).$$

Usually,

$$Q_n(\theta) = n^{-1} \sum_{i=1}^{n} \psi(z_i, \theta),$$

but we will not force this structure at first. Note that GMM estimators are not contained in the average framework.

That is to say, our minimization estimator is assumed to minimize a random function $Q_n(\theta)$, and we will not be explicit about the exact form of $Q_n(\cdot)$. The $\Theta$ is the parameter space, the space over which the objective function is minimized.

Next, we need to acquire an understanding of a key aspect of the consistency proof: the uniform law of large numbers.

5.1 The uniform law of large numbers

From the law of large numbers, we expect that

$$n^{-1} \sum_{i=1}^{n} (\psi(z_i, \theta) - E\psi(z_i, \theta)) \xrightarrow{\text{as}} 0$$

for all $\theta \in \Theta$, where $\Theta$ denotes some parameter space (that in most cases is preferably chosen as $\mathbb{R}^k$). Therefore, one might suspect that the condition

$$\sup_{\theta \in \Theta} |n^{-1} \sum_{i=1}^{n} (\psi(z_i, \theta) - E\psi(z_i, \theta))| \xrightarrow{\text{as}} 0.$$ 

is (nearly) always satisfied. This is not true, however; there is a difference between the pointwise LLN and the uniform LLN. As an example of the problem involved, consider $\Theta = [0, 1]$ and

$$f_n(\theta) = I(0 < \theta \leq 1/n).$$
Then, for all $\theta$, $\lim_{n \to \infty} f_n(\theta) = 0$, while still $\sup_{\theta \in \Theta} f_n(\theta) = 1$ for all $n$. One might argue that this example is nonstochastic and therefore probably not completely satisfactory. The space $\theta = [0, 1]$ is compact however, and the $f_n(\cdot)$ function can also be modified to such as to give an example of a continuous function with similar behavior. If stochastic convergence is introduced, complications can only multiply.

Therefore, the conclusion has to be that uniform convergence and pointwise convergence are two different concepts. Obviously, uniform convergence implies pointwise convergence.

### 5.2 Consistency: standard assumptions

For consistency of general minimization estimators, the following assumptions suffice. Notice that the assumption of compactness of $\Theta$ has not (yet) been made:

**Assumption 1**

1. $Q_n(\theta)$ and $\hat{\theta}_n = \arg\min_{\theta \in \Theta} Q_n(\cdot)$ are a well-behaved random functions of $\theta$.

2. Uniform convergence: $\sup_{\theta \in \Theta} |Q_n(\theta) - Q(\theta)| \overset{a.s.}{\to} 0$, for a deterministic function $Q(\cdot)$ and $Q(\cdot)$ is continuous on $\Theta$.

3. Identification (part 1): there exists a unique $\theta^*$ such that $Q(\theta^*) = \min_{\theta \in \Theta} Q(\theta)$.

4. Identification (part 2): For all $\delta > 0$, $\inf_{\theta \in \Theta; |\theta - \theta^*| > \delta} Q(\theta) > Q(\theta^*)$.

Assumption 1.1 defines $\hat{\theta}_n$ and rules out technical problems involving the existence of random variables (it rules out measurability problems). Assumptions 1.3 and 1.4 are regularity assumptions that need to ensure that there is one single minimizer of the limit objective function that is well-behaved. These conditions are identifiability conditions and will fail to hold, for example, if we have multicollinearity in our regressors. Note that in the case of exact multicollinearity, $Q_n(\cdot)$ does not have a unique minimizer, while the condition above says that the limit objective function should not have multiple minimizers. Assumption 1.4 holds automatically if $\Theta$ is compact and 1.3 holds. Assumption 1.2 is a uniform law of large numbers.

Using the above assumptions, the following theorem can be established:

**Theorem 3** Assume $\hat{\theta}_n$ exists as a random variable for all $n$. Then Assumptions 1.1, 1.2 and 1.3 imply that $\hat{\theta}_n \overset{a.s.}{\to} \theta^*$ almost surely.
**Proof:** We will show that, for \( n \) large enough, \( \hat{\theta}_n \) cannot be further from \( \theta^* \) than \( \delta \), say. The result then follows since \( \delta \) was arbitrary. The idea of the proof is to show that

\[
Q_n(\theta) - Q_n(\theta^*)
\]

exceeds zero if \( |\theta - \theta^*| > \delta \) and \( n \) is large, and therefore, for \( n \) large enough, \( |\hat{\theta}_n - \theta^*| \leq \delta \). We show this by defining \( \varepsilon > 0 \) as

\[
\inf_{|\theta - \theta^*| > \delta} Q(\theta) = Q(\theta^*) + \varepsilon
\]

and noting that for \( n \) large enough

\[
\sup_{\theta \in \Theta} |Q_n(\theta) - Q(\theta)| < \varepsilon/2.
\]

Therefore,

\[
\inf_{|\theta - \theta^*| > \delta} Q_n(\theta) - Q_n(\theta^*)
\]

\[
= \inf_{|\theta - \theta^*| > \delta} Q_n(\theta) - Q(\theta) + Q(\theta)
\]

\[
+ Q_n(\theta^*) - Q(\theta^*) + Q(\theta^*)
\]

\[
\geq \inf_{|\theta - \theta^*| > \delta} Q(\theta) - Q(\theta^*) - \sup_{\theta \in \Theta} |Q_n(\theta) - Q(\theta)|
\]

\[
- |Q_n(\theta^*) - Q(\theta^*)|
\]

\[
> \varepsilon - \varepsilon/2 - \varepsilon/2 = 0.
\]

Therefore, for \( n \) large enough, \( |\hat{\theta}_n - \theta^*| \leq \delta \) since the criterion function is too large for other values of \( \hat{\theta}_n \). Therefore, \( \lim_{n \to \infty} \hat{\theta}_n = \theta^* \) almost surely.

An alternative line of reasoning is as follows. Note that, since \( \theta^* \) minimizes \( E\psi(z_i, \theta) \),

\[
0 \leq Q(\hat{\theta}_n) - Q(\theta^*).
\]

Next, note that

\[
0 \leq Q(\hat{\theta}_n) - Q(\theta^*)
\]

\[
= Q(\hat{\theta}_n) - Q_n(\hat{\theta}_n) + Q_n(\hat{\theta}_n) - Q(\theta^*)
\]

\[
\leq Q(\hat{\theta}_n) - Q_n(\hat{\theta}_n) + Q_n(\theta^*) - Q(\theta^*)
\]

23
\[ \leq 2 \sup_{\theta \in \Theta} |Q_n(\theta) - Q(\theta)| \xrightarrow{as} 0. \]

The second inequality follows because \( \hat{\theta}_n \) minimizes \( Q_n(\theta) \) over \( \theta \in \Theta \). So,

\[ \lim_{n \to \infty} Q(\hat{\theta}_n) = Q(\theta^*), \]

and therefore by the uniqueness assumptions of Assumption 1 and Assumption 2 that ensure that \( \hat{\theta}_n \) will not “wonder off”,

\[ \hat{\theta}_n \xrightarrow{as} \theta^*. \]

### 5.3 The uniform convergence assumption

Consider the following example. If we set

\[ Q_n(\theta) = n^{-1} \sum_{i=1}^{n} (y_i - \theta)^2, \]

then

\[ Q(\theta) = E(y_i - \theta)^2 \]

and

\[ Q_n(\theta) - Q(\theta) = n^{-1} \sum_{i=1}^{n} (y_i^2 - Ey_i^2) - 2\theta n^{-1} \sum_{i=1}^{n} (y_i - Ey_i). \]

This implies that

\[ \sup_{\theta \in \mathbb{R}} |Q_n(\theta) - Q(\theta)| = \infty, \]

and therefore the uniform convergence condition fails if we set \( \Theta = \mathbb{R} \). If \( \Theta = [-1, 1] \), say, the supremum will be attained at either \( \theta = -1 \) or \( \theta = 1 \), and uniform convergence will hold under standard regularity conditions.

As it turns out, the uniform convergence condition is nonlinear settings is easy to show under some conditions if the parameter space is compact. It is not clear what this implies for empirical practice. Formally, the standard arguments as will be explained below only hold if the researcher promises to never let the parameter outside of an a priori identified parameter space.
It is possible to relax the compactness assumption in the case of a globally convex or concave objective function (such as in the case of least squares and ML estimation of probit, logit, and (after reparametrizing) Tobit models.

In general however, standard theory will force us to make the compactness assumption. It is possible to show examples of inconsistent estimators in cases where all the assumptions above hold, except that uniform convergence is weakened to pointwise convergence.