Topic 11

Asymptotic and Large Sample Results

1 What is in this document?

This document contains two separate sets of extended notes. One set is from Peter Kennedy's book (and this appears before our main lecture slides for Topic 11), and the other set is from some old lectures by DrVH (and this appears after our main lecture slides for Topic 11).

Extended notes that have been provided below BEFORE the lecture slides...

For those who need a reasonably non-technical primer on asymptotics, and/or how to think about asymptotics in the linear regression context, DrRS can recommend "Appendix C" reproduced below for you from Peter Kennedy's book. Kennedy's Appendix C introduces the big picture behind Topic 11, and it also recaps some of the key underlying theoretical concepts needed, such as those of convergence in probability and in distribution. It also explains the logic behind some of the steps you will encounter in your consistency and asymptotic normality proofs.

Note that the excerpt from Kennedy does not cover the entirety of the content that DrVH will teach, particularly in relation to non-linear estimators and approximate SEVs. Nevertheless, it does provide nice explanations of the more basic ideas. Hence, we include it here in these extended notes.

Extended notes that have been provided below AFTER the lecture slides...

Of course, some of the material that is discussed by DrVH is more technical than the coverage in Kennedy's book. For supplemental background reading on more technical areas, please see the extended notes that have been presented below AFTER the main lecture slides appear. These are from DrVH directly.

Appendix C

A Primer on Asymptotics

The rationale behind asymptotic distribution theory and the reasons for econometricians' interest in it are presented in chapter 2. In essence, an asymptotic analysis is a means of obtaining information enabling us better to understand finite sample distributions and so to produce good approximations. The purpose of this appendix is to provide an overview of the technical details of asymptotics. Readers are warned that to keep this presentation readable, many not-quite-correct statements appear; for those interested in mastering the details, several recent advanced textbooks have good presentations, for example, Greene (2008, pp. 63–75 and appendix D) and Judge *et al.* (1985, chapter 5). A good advanced reference is Greenberg and Webster (1983, chapter 1). White (1984) is very advanced, Kmenta (1986, pp. 163–72) and Darnell (1994, pp. 45–9, 290–3, 217–22) have good expositions at the beginner level.

Asymptotic distribution theory is concerned with what happens to a statistic, say $\hat{\beta}$, as the sample size N becomes very large. To emphasize the role of the sample size, $\hat{\beta}$ is sometimes written as $\hat{\beta}_N$. In particular, interest focuses on two things:

(a) Does the distribution of $\hat{\beta}_N$ collapse on a particular value (i.e., become heavily concentrated in the neighborhood of that value) as the sample size becomes very large? This leads to the large-sample concept of consistency.

(b) Does the distribution of $\hat{\beta}_N$ approximate a known form (e.g., the normal distribution) as the sample size becomes very large? This allows the development of large-sample hypothesis testing procedures.

To address these questions, two concepts of convergence are employed. Convergence in probability is used for (a) above, and convergence in distribution is used for (b).

1 Convergence in Probability

Suppose that as the sample size becomes very large the distribution of $\hat{\beta}_N$ collapses on the value k. Then $\hat{\beta}_N$ is said to *converge in probability* to k, or has *probability limit* k, written as plim $\hat{\beta} = k$. If k equals β, the number that $\hat{\beta}$ is estimating, $\hat{\beta}$ is said to be *consistent*; $k - \beta$ is called the *asymptotic bias* of $\hat{\beta}$ as an estimator of β.

A popular means of showing consistency is to show that the bias and the variance of $\hat{\beta}_N$ both approach zero as the sample size becomes very large. This is called *convergence in quadratic mean* or *convergence in mean square*; it is a sufficient condition for convergence in probability. Consider, for example, the sample mean statistic from a sample drawn randomly from a distribution with mean μ and variance σ^2. Because the sample mean is unbiased in small samples, it has zero bias also in large samples, and because its variance is σ^2/N, its variance approaches zero as the sample size becomes very large. Thus the sample mean converges in quadratic mean and therefore is a consistent estimator of μ.

A major reason for using asymptotic distribution theory is that the algebra associated with finding (small-sample) expected values can become formidable whenever nonlinearities are involved. In particular, the expected value of a nonlinear function of $\hat{\beta}$ say, is not equal to the nonlinear function of the expected value of $\hat{\beta}$. Why this happens is explained in the technical notes to section 2.8. This problem disappears when using asymptotics, however, because the plim of a nonlinear (continuous) function of $\hat{\beta}$ is the nonlinear function of the plim of $\hat{\beta}$. This is referred to as *Slutsky's theorem*; the reason for this is also explained in the technical notes to section 2.8. As an example, suppose you have an unbiased estimator β^* of the multiplier $\pi = 1/(1 - \beta)$ but you wish to estimate β. Now $\beta = 1 - \pi^{-1}$ so it is natural to suggest using $1 - (\pi^*)^{-1}$ to estimate β. Since this is a nonlinear function of the unbiased estimate β^* it will be biased, but, thanks to Slutsky's theorem, asymptotically unbiased.

Consider now the task of showing the consistency of the ordinary least squares (OLS) estimator in the classical linear regression (CLR) model $y = X\beta + \varepsilon$. Since β^{OLS} can be written as $\beta + (X'X)^{-1}X'\varepsilon$ we have plim $\beta^{\text{OLS}} = \beta + \text{plim}(X'X)^{-1}X'\varepsilon$. It is instructive to spell out fully the logic of the remainder of the argument.

(a) $(X'X)^{-1}X'\varepsilon$ is multiplied and divided by N, producing $(X'X/N)^{-1}(X'\varepsilon/N)$.

(b) Slutsky's theorem is used to break the plim into two halves, namely

$$\text{plim}(X'X)^{-1}X'\varepsilon = \text{plim}(X'X/N)^{-1}\text{plim}(X'\varepsilon/N)$$

and then is used again to bring the first plim inside the inverse sign, producing

$$[\text{plim}(X'X/N)]^{-1}\text{plim}(X'\varepsilon/N).$$

(c) It should now be evident why the Ns were inserted. $X'X$ is a matrix consisting of sums, with each extra observation adding something to each of these sums. As N becomes very large some of these sums will undoubtedly become infinite. (All the diagonal elements are sums of squares; if there is an intercept, the upper left corner of this matrix is equal to N.) Consequently, it would not make much sense to find plim$(X'X)$. In contrast, by examining plim$(X'X/N)$ we are in effect looking at the average values of the elements of the $X'X$ matrix, and these are finite under a fairly broad set of assumptions as the sample size becomes very large.

(d) To proceed further, it is necessary to make some assumption about how extra observations on the independent variables are obtained as the sample size grows. The standard assumption made is that these extra observations are such that plim$(X'X/N)$ is equal to a finite, invertible matrix Q. Loosely speaking, Q can be thought of as the expected value of the $X'X$ matrix for a sample of size one. Theoretical results are often expressed in terms of Q; it must be remembered that, operationally, Q will be estimated by $X'X/N$.

(e) We now have that plim $\beta^{\text{OLS}} = \beta + Q^{-1}\text{plim}(X'\varepsilon/N)$. It is tempting at this stage to use Slutsky's theorem once again to break plim$(X'\varepsilon)$ into plimX'plimε. This would not make sense, however. Both X and ε have dimension N, therefore as the sample size grows X becomes a bigger and bigger matrix and ε a longer and longer vector.

(f) What does a typical element of $X'\varepsilon/N$ look like? Suppose the jth explanatory variable is w. Then the jth element of $X'\varepsilon/N$ is $\Sigma w_i\varepsilon_i/N$. In the CLR model, the w is fixed in repeated samples, and the expected value of ε is zero, so the expected value of $\Sigma w_i\varepsilon_i/N$ is zero. What about the variance of $\Sigma w_i\varepsilon_i/N$? It is equal to $\sigma^2\Sigma w_i^2/N^2 = (\sigma^2/N)\Sigma w_i^2/N$, which approaches zero as the sample size becomes very large (since the term $\Sigma w_i^2/N$ is finite, approaching the jth diagonal element of Q). Thus because the expected value and variance both approach zero as the sample size becomes very large (i.e., convergence in quadratic mean), the plim of $X'\varepsilon/N$ is the zero vector, and thus plim $\beta^{\text{OLS}} = \beta$; β^{OLS} is consistent in the CLR model.

(g) A more straightforward way of obtaining convergence in quadratic mean for this case, and thus consistency, is to note that because β^{OLS} is unbiased in small samples it is also unbiased in large samples, and that the variance of β^{OLS} can be written as $\sigma^2(X'X)^{-1} = (\sigma^2/N)(X'X/N)^{-1}$, which approaches zero as the sample size becomes very large.

(h) The observant reader would have noticed that the assumption that plim$(X'X/N)$ equals a finite invertible matrix rules out a very common case, namely a regressor following a growth trend. If the regressor values grow as the sample size grows, plim$(X'X/N)$ will become infinite. Fortunately, this does not cause insurmountable problems, mainly because if this becomes infinite its inverse becomes zero. Look again at the argument in (g) above. The key is that $(\sigma^2/N)(X'X/N)^{-1}$ converges to zero as the sample size becomes very large; this comes about because (σ^2/N) approaches zero while $(X'X/N)^{-1}$ is assumed to approach a finite value. In the case of a trending regressor this latter term also approaches zero, aiding the convergence of $(\sigma^2/N)(X'X/N)^{-1}$ to zero.

(i) A key element in (f) above is that the expected value of $\Sigma w_t\varepsilon_t/N$ is zero. If w is stochastic, rather than fixed in repeated samples, this will happen if w is contemporaneously independent of the error term. This reveals why it is only contemporaneous dependence between a regressor and the error term that leads to asymptotic bias.

2 Convergence in Distribution

Suppose that as the sample size becomes very large the distribution f_N of $\hat{\beta}_N$ becomes virtually identical to a specific distribution f. Then $\hat{\beta}_N$ is said to *converge in distribution* to f (sometimes expressed as converging in distribution to a variable whose distribution is f). The distribution f is called the *limiting distribution* of $\hat{\beta}_N$; the intention is to use this limiting distribution as an approximation for the unknown (or intractable) small-sample distribution of $\hat{\beta}_N$. Two difficulties are apparent.

First, we saw earlier that in most applications the distribution of $\hat{\beta}_N$ collapses to a spike, so it does not make sense to use it to approximate the small-sample distribution of $\hat{\beta}_N$. This difficulty is overcome by transforming/normalizing $\hat{\beta}_N$ to prevent its distribution from collapsing. The most common way of accomplishing this is to focus attention on the distribution of $\sqrt{N}(\hat{\beta}_N - \beta)$. For the example of β^{OLS} in the CLR model, it is easy to see that as the sample size becomes very large the mean of $\sqrt{N}(\beta^{\text{OLS}} - \beta)$ is zero and its variance is σ^2Q^{-1}.

Second, how are we going to know what form (e.g., normal distribution) the distribution of $\sqrt{N}(\hat{\beta}_N - \beta)$ takes as the sample size becomes very large? This problem is solved by appealing to a *central limit theorem.* Central limit theorems in effect say that the sample mean statistic is distributed normally when the sample size becomes very large, that is, that the limiting distribution of $\sqrt{N}(\hat{\theta}_N - \theta)$ is a normal distribution if $\hat{\theta}_N$ is a sample average. It is remarkable that so many statistics can be shown to be functions of a sample mean statistic, allowing a central limit theorem to be exploited to derive limiting distributions of known form.

To illustrate this consider once again β^{OLS} in the CLR model. If the errors were distributed normally, β^{OLS} would be normally distributed with mean β and variance $\sigma^2(X'X)^{-1}$. If the errors are not distributed normally, the distribution of β^{OLS} is difficult to describe and to utilize for hypothesis testing. Instead of trying to derive the exact distribution of β^{OLS} in this circumstance, what is usually done is to approximate this exact distribution with what is called the *asymptotic distribution* of β^{OLS}.

3 Asymptotic Distributions

The first step in finding this asymptotic distribution is to find the limiting distribution of $\sqrt{N}(\beta_N^{\text{OLS}} - \beta) = (X'X/N)^{-1}(X'\varepsilon/\sqrt{N})$. Look first at $(X'\varepsilon/\sqrt{N})$, which can be rewritten as $\sqrt{N}(X'\varepsilon/N)$. Following our earlier discussion, suppose the jth explanatory variable is w. Then the jth element of $\sqrt{N}(X'\varepsilon/N)$ is $\sqrt{N}(\Sigma w_t\varepsilon_t/N)$. Notice that $\Sigma w_t\varepsilon_t/N$ is a sample average of the $w_t\varepsilon_t$s, and that the common mean of the $w_t\varepsilon_t$s is zero. Consequently, a central limit theorem can be applied to show that the limiting distribution of $\sqrt{N}(X'\varepsilon/N)$ is normal with mean zero. The variance can be derived as σ^2Q.

We can now apply a very useful theorem concerning the interaction between plims and limiting distributions: if one variable has a plim and another variable has a limiting distribution, then when dealing with their product the first variable can be treated as a constant in so far as the limiting distribution of that product is concerned. Thus, for example, suppose plim $a_N = a$ and the limiting distribution of b_N is normal with mean μ and variance σ^2. Then the limiting distribution of $a_N b_N$ is normal with mean $a\mu$ and variance $a^2\sigma^2$. To be even more specific, suppose $\sqrt{N}(\bar{x} - \mu)$ has limiting distribution $N(0, \sigma^2)$, and plim $s^2 = \sigma^2$; then the limiting distribution of $\sqrt{N}(\bar{x} - \mu)/s = (\bar{x} - \mu)/(s/\sqrt{N})$ is $N(0,1)$.

We wish to use this theorem to find the limiting distribution of $\sqrt{N}(\beta_N^{\mathrm{OLS}} - \beta) = (X'X/N)^{-1}(X'\varepsilon/\sqrt{N})$. Since $\mathrm{plim}(X'X/N)^{-1} = Q^{-1}$ and the limiting distribution of $(X'\varepsilon/\sqrt{N})$ is $N(0, \sigma^2 Q)$, the limiting distribution of $(X'X/N)^{-1}(X'\varepsilon/\sqrt{N})$ is $N(0, Q^{-1}\sigma^2 QQ^{-1}) = N(0, \sigma^2 Q^{-1})$.

It is customary, although not technically correct, to use the expression "the asymptotic distribution of β^{OLS} is $N(\beta, (\sigma^2/N)Q^{-1})$" to refer to this result. This distribution is used as an approximation to the unknown (or intractable) small-sample distribution of β^{OLS}; in this example, β^{OLS} is said to be *asymptotically normally distributed* with mean β and asymptotic variance $(\sigma^2/N)Q^{-1}$. On the assumption that the sample size is large enough for this distribution to be a good approximation (it is remarkable that such approximations are typically quite accurate for samples of modest size), hypothesis testing proceeds in the usual fashion, in spite of the errors not being normally distributed. Since Q is estimated by $X'X/N$, operationally the variance $(\sigma^2/N)Q^{-1}$ is estimated by the familiar $s^2(X'X)^{-1}$.

Joint hypotheses are tested via the usual $F(J, N-K)$ statistic, or, in its asymptotic incarnation, J times this F statistic, which is distributed asymptotically as $\chi^2(J)$. This is justified by appealing to another extremely useful theorem: if a statistic converges in distribution to x, then a continuous function g of that statistic converges in distribution to $g(x)$. For example, if the limiting distribution of θ^* is $N(0, 1)$, then the limiting distribution of $(\theta^*)^2$ is $\chi^2(1)$.

Another way of dealing with nonlinearities is to appeal to the result that if a statistic $\hat{\beta}$ is distributed asymptotically normally then a continuous function g of that statistic is distributed asymptotically normally with mean $g(\mathrm{plim}\ \hat{\beta})$ and variance equal to the variance of $\hat{\beta}$ times the square of the first derivative of g with respect to $\hat{\beta}$, as described in appendix B and the technical notes to chapter 2.

Notes

- The terminology $\xrightarrow{p}$ is used to express convergence in probability, so $a_N \xrightarrow{p} a$ means that a_N converges in probability to a. The terminology $\xrightarrow{d}$ is used to express convergence in distribution, so $\sqrt{N}(a_N - a) \xrightarrow{d} x$ means that the limiting distribution of a_N is the distribution of x. If the distribution of x is known to be $N(0, 1)$, for example, this is often written as $\sqrt{N}(a_N - a) \xrightarrow{d} N(0, 1)$.
- A formal definition of consistency is as follows: an estimator $\hat{\beta}$ of β is consistent if the probability that $\hat{\beta}$ differs in absolute value from β by less than some preassigned positive number δ (however small) can be made as close to one as desired by choosing a suitably large-sample size. This is usually written as

$$\mathrm{plim}\ \hat{\beta} = k \qquad \text{if } \lim_{N\to\infty} \mathrm{prob}\ (|\hat{\beta} - k| < \delta) = 1$$

where δ is any arbitrarily small positive number.

- The discussion above has on occasion referred to the plim as the *asymptotic expectation.* Unfortunately, there is some confusion in the literature concerning this: some people define the asymptotic expectation to be the plim, but most define it to be the limit of the expected value, which is not the same thing. Although in virtually all practical applications the two are identical, which explains why most people treat them as being equivalent, it is possible to find cases in which they differ. It is instructive to look at some examples.

1. Suppose $\mathrm{prob}(\hat{\beta} = \beta) = 1 - 1/N$ and $\mathrm{prob}(\hat{\beta} = N) = 1/N$ where N is the sample size. The plim is β, but the asymptotic expectation is $\beta + 1$.
2. Suppose we have a sample on x of size N and estimate the population mean μ by $\mu^* = x_1/2 + \Sigma x_i/2(N-1)$ where the summation runs from 2 to N. The asymptotic expectation is μ, but the plim is $x_1/2 + \mu/2$.
3. Consider the inverse of the sample mean statistic as an estimate of a nonzero population

mean μ. Its plim is μ^{-1}, but its asymptotic expectation does not exist (because of the possibility that the sample mean is zero).

The plim and the asymptotic expectation will be the same whenever they both exist and the variance goes to zero as the sample size goes to infinity. In general, refer to consistency and plims; avoid using asymptotic expectation.

- The above examples illustrate why convergence in quadratic mean is not a necessary condition for consistency.
- A stronger form of convergence in probability, called *almost sure convergence*, is sometimes encountered. The former allows some erratic behavior in the converging sequence, whereas the latter does not.
- The order of a statistic is sometimes encountered when dealing with asymptotics. A statistic θ^* is said to be at most of order N^k if plim θ^*/N^k is a nonzero constant. For example, since $X'X/N$ converges to Q, $X'X$ is at most of order N. The big O notation, $X'X = O(N)$ is used to denote this. The little o notation $\theta = o(N^k)$ means the statistic θ is of smaller order than N^k, implying that plim $\theta/N^k = 0$. Typically, for coefficient estimators that are biased, but consistent, the order of their bias is $1/\sqrt{N}$, meaning that this bias disappears at a rate proportional to the square root of the sample size: plim $\sqrt{N}$(bias) is a constant. For the OLS estimator of the cointegrating vector, discussed in chapter 19, the bias disappears at a rate proportional to N, explaining why this estimator is called "superconsistent." In this example, the usual transformation $\sqrt{N}(\hat{\beta}_N - \beta)$ suggested earlier is inappropriate; the transformation $N(\hat{\beta}_N - \beta)$ must be used for this case.
- Note that although strictly speaking the limiting/asymptotic distribution of β^{OLS} is a degenerate spike at β, many econometricians speak of the asymptotic distribution of β^{OLS} as normal with mean β and asymptotic variance $(\sigma^2/N)Q^{-1}$. This should be interpreted as meaning that the limiting distribution of $\sqrt{N}(\beta^{OLS} - \beta)$ is normal with mean zero and variance $\sigma^2 Q^{-1}$.
- Continuing to speak loosely, an estimator's asymptotic variance cannot be calculated by taking the limit of that estimator's variance as the sample size becomes very large, because usually that limit is zero. In practice, it is common for it to be calculated as

$$\text{Asy. Var}\,\hat{\beta} = (1/N) \lim_{N\to\infty} NV(\hat{\beta}).$$

- There exist several central limit theorems, which are applicable in differing circumstances. Their general flavor is captured by the following: if $\bar{x}$ is the average of N random drawings from probability distributions with common finite mean μ and (differing) finite variances, then the limiting distribution of $\sqrt{N}(\bar{x} - \mu)$ is normal with mean zero and variance the limit of the average of the variances.
- A consistent estimator is said to be *asymptotically efficient* if its asymptotic variance is smaller than the asymptotic variance of any other consistent estimator. Sometimes this refers only to estimators that are distributed asymptotically normally. The maximum likelihood estimator, whose asymptotic variance is given by the Cramer–Rao lower bound, is asymptotically efficient, and therefore is used as a benchmark in this regard.

Topic 11 Header Slide

- Topic 11. Asymptotic and large sample results

  Asymptotic results: refer to theoretical results (about the probabilistic behaviour of our estimator) that hold only in the limit as S passes to ∞.

  Large sample results: refer to asymptotic results that are thought to hold approximately for sufficiently large (albeit finite) S.

Estimator	Notation
$1.$	$\hat{\beta}_{OLS}$
$2.$	$\hat{\beta}_{LAD}$
$3.$	$\hat{\beta}_{Lstar}$
$4.$	$\hat{\beta}_{GMM}$
$5a.$	$\hat{\beta}_{IGLS}$
$5b.$	$\hat{\beta}_{FGLS}$
$6.$	$\hat{\beta}_{MLE}$

For any method, generically denoted $\hat{\theta}_{method}$, we define

$$SEV(\hat{\theta}_{method}) = \hat{\theta}_{method} - \theta^{true}.$$

Common form of SEV for analytic methods

Consider analytic methods for estimation (i.e., whereby the objective function is twice continuously differentiable). We summarise the common structure/form of the SEV for such estimators:

- **Summary 1.** For analytic methods that are linear in y and ε^{true}, there exists a $k \times k$ matrix, B_s, and a $k \times 1$ vector, a_s, such that

$$SEV(\hat{\beta}_{method}) = \hat{\beta}_{method} - \beta^{true} = \left(\sum_{s=1}^{S} B_s\right)^{-1} \sum_{s=1}^{S} a_s.$$

  Example: We saw, for OLS, the definitions: $B_s = x_s x_s'$, and $a_s = x_s \varepsilon^{true}$.

- **Summary 2.** For analytic methods that are non-linear, there exists a $k \times k$ matrix, B_s, and a $k \times 1$ vector, a_s, such that

$$SEV(\hat{\beta}_{method}) = \hat{\beta}_{method} - \beta^{true} \approx \left(\sum_{s=1}^{S} B_s\right)^{-1} \sum_{s=1}^{S} a_s,$$

  where the approximation ("$\approx$") is reasonable for sufficiently large S, and the approximation in fact becomes exact in the limit as S passes to ∞.
  Example: We will see, for MLE in the general linear/non-linear case, the definitions: $B_s = -\ell_s^{\beta\beta'}(\beta^{true})$, and $a_s = \ell_s^{\beta}(\beta^{true})$, where ℓ_s^{β} denotes the score contribution (vector) by the s-th observation and $\ell_s^{\beta\beta'}$ denotes the corresponding second-order derivative (matrix).

Sample averages and normalised sample averages (1 of 2)

In each case above (linear/non-linear), we consider algebraic structures involving sample averages:

- **Summary 1.** For analytic methods that are linear in y and ε^{true}, there exists a $k \times k$ matrix, B_s, and a $k \times 1$ vector, a_s, such that

$$SEV(\hat{\beta}_{method}) = \hat{\beta}_{method} - \beta^{true} = \left(\frac{1}{S} \sum_{s=1}^{S} B_s \right)^{-1} \frac{1}{S} \sum_{s=1}^{S} a_s.$$

  Example: We saw, for OLS, the definitions: $B_s = x_s x_s'$, and $a_s = x_s \varepsilon^{true}$.

Above, by scaling throughout by $(1/S)$, we compute sample averages in both the inverse and non-inverse term of the SEV. This formulation of the SEV will be extremely useful to us for consistency proofs (convergence in probability).

In contrast, for asymptotic normality proofs (convergence in distribution), it is useful to consider the SEV in terms of normalised sample averages as follows:

$$\sqrt{S} SEV(\hat{\beta}_{method}) = \sqrt{S}(\hat{\beta}_{method} - \beta^{true}) = \left(\frac{1}{S} \sum_{s=1}^{S} B_s \right)^{-1} \frac{1}{\sqrt{S}} \sum_{s=1}^{S} a_s.$$

Sample averages and normalised sample averages (2 of 2)

- **Summary 2.** For analytic methods that are non-linear, there exists a $k \times k$ matrix, B_s, and a $k \times 1$ vector, a_s, such that

$$SEV(\hat{\beta}_{method}) = \hat{\beta}_{method} - \beta^{true} \approx \left(\frac{1}{S} \sum_{s=1}^{S} B_s \right)^{-1} \frac{1}{S} \sum_{s=1}^{S} a_s,$$

  where the approximation ("$\approx$") is reasonable for sufficiently large S, and the approximation in fact becomes exact in the limit as S passes to ∞.

  Example: We will see, for MLE in the general linear/non-linear case, the definitions: $B_s = -\ell_s^{\beta\beta'}(\beta^{true})$, and $a_s = \ell_s^{\beta}(\beta^{true})$, where ℓ_s^{β} denotes the score contribution (vector) by the s-th observation and $\ell_s^{\beta\beta'}$ denotes the corresponding second-order derivative (matrix).

Above, by scaling throughout by $(1/S)$, we compute sample averages in both the inverse and non-inverse term of the approximate SEV. This formulation of the approximate SEV will be extremely useful to us for consistency proofs (convergence in probability).

In contrast, for asymptotic normality proofs (convergence in distribution), it is useful to consider the approximate SEV in terms of normalised sample averages as follows:

$$\sqrt{S}(SEV(\hat{\beta}_{method})) = \sqrt{S}(\hat{\beta}_{method} - \beta^{true}) \approx \left(\frac{1}{S} \sum_{s=1}^{S} B_s \right)^{-1} \frac{1}{\sqrt{S}} \sum_{s=1}^{S} a_s.$$

Building blocks of the SEV

Our interest, broadly speaking, is in analysing what happens to $SEV(\hat{\beta}_{method})$ as $S \to \infty$. We will do so by considering each of the following sample averages or normalised sample averages:

$$\frac{1}{S}\sum_{s=1}^{S} a_s,$$

$$\frac{1}{S}\sum_{s=1}^{S} B_s,$$

$$\frac{1}{\sqrt{S}}\sum_{s=1}^{S} a_s,$$

$$\left(\frac{1}{S}\sum_{s=1}^{S} B_s\right)^{-1},$$

$$\left(\frac{1}{S}\sum_{s=1}^{S} B_s\right)^{-1} \frac{1}{S}\sum_{s=1}^{S} a_s = SEV(\hat{\beta}_{method}),$$

$$\left(\frac{1}{S}\sum_{s=1}^{S} B_s\right)^{-1} \frac{1}{\sqrt{S}}\sum_{s=1}^{S} a_s = \sqrt{S}(SEV(\hat{\beta}_{method})).$$

Extremely important remarks about the previous slide

- **Remark 1.** Recall that $\frac{1}{S}\sum_{s=1}^{S}(\cdot)$ is the first sample moment (or sample average).
- **Remark 2.** We define $\frac{1}{\sqrt{S}}\sum_{s=1}^{S}(\cdot)$ to be the first normalised sample moment (or normalised sample average).
- **Remark 3.** Inverses of matrices (unless singular) and products of matrices (unless undefined) are examples of continuous functions (or mappings) of their arguments.
- **Remark 4.** We will evaluate the behaviour of first sample moments using laws of large numbers (LLNs).
- **Remark 5.** We will evaluate the behaviour of first normalised sample moments using the central limit theorem (CLT).
- **Remark 6.** We will evaluate the behaviour of continuous functions (or mappings) of sample averages using Slutsky's theorem (i.e., our first continuous mapping theorem, CMT1).
- **Remark 7.** We will evaluate the behaviour of continuous functions (or mappings) of sample averages and normalised sample averages using Cramér's theorem (i.e., our second continuous mapping theorem, CMT2).

Specific SEVs for various analytic methods

- For method 1 (OLS) and method 4 (GMM with $A2linear$), we have

$$SEV(\hat{\beta}_{OLS}) = \hat{\beta}_{OLS} - \beta^{true} = (X'X)^{-1}X'\varepsilon^{true} = \left(\sum_{s=1}^{S} x_s x_s'\right)^{-1} \sum_{s=1}^{S} x_s \varepsilon_s^{true}.$$

- Method 2 (LAD) is not analytic; and the SEV for method 3 (Lstar) is not useful to consider.

- For method 5a (IGLS), suppose (for convenience) that we have $A4\Omega$ with a diagonal Ω. Then, we have

$$\begin{aligned} SEV(\hat{\beta}_{IGLS}) &= \hat{\beta}_{IGLS} - \beta^{true} = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}\varepsilon^{true} \\ &= \left(\sum_{s=1}^{S} x_s \omega_{ss} x_s'\right)^{-1} \sum_{s=1}^{S} x_s \omega_{ss} \varepsilon_s^{true}, \end{aligned}$$

  where weight $\omega_{st} = \left[\Omega^{-1}\right]_{st}$ is the (s,t)-th element of Ω^{-1}, for $s, t = 1, ..., S$.

  The exact weights, ω_{st}, are unimportant; what is important is that the SEV of the IGLS estimator can be expressed in the same common form as other analytic estimators.

- We consider method 6 (MLE) in a lot of detail in the next slides.

Specific SEV for MLE (whether linear/non-linear in y)

Recall that for a parametric estimation problem for p-dimensional parameter vector θ^{true}, in the generalised LRM, via maximum likelihood estimation, we impose the assumptions $A1, A2linear, \geq A3Rmi, A4\Omega.independent$, and some $A5specific$.

In the previous sentence, $A4\Omega.independent$ refers to the $A4\Omega$ assumption with the addition of independence imposed across the s dimension (so that Ω is necessarily diagonal). Recall that this assumption makes MLE more manageable since we can thereby obtain the overall likelihood as the product of the marginal contributions to the overall likelihood by each observation.

Under the given specification, and denoting $\{y, X\}$ as $data$, we have

$$\hat{\theta}_{MLE} = \arg\max \ell(\theta; data) = \arg\max \sum_{s=1}^{S} \ell_s(\theta; data),$$

where $\ell_s(\theta; data) = \log f_s(y_s|X; \theta)$ for $s = 1, ..., S$, are the marginal contributions of each observation to the overall log likelihood.

We need to maximise an objective function but FOCs/SOCs can only be defined if the likelihood function is twice continuously differentiable. In the absence of twice continuous differentiability – e.g., under $A5LAD$ or $A5LDE$ – no analytic solution to the maximisation problem exists.

Let us suppose, for the moment, that the likelihood does admit a continuous second derivative. Now consider the maximisation procedure as outlined in the following slides.

FOC and SOC for MLE (whether linear/non-linear in y)

Under twice continuous differentiability of the likelihood function, and under the given model specification (on the previous slide), the ML estimator is (at least implicitly) defined by the following first and second order conditions (FOCs/SOCs):

FOC:

$$\frac{\partial \ell(\theta; data)}{\partial \theta}\Big|_{\theta=\hat{\theta}_{MLE}} = \sum_{s=1}^{S} \ell_s^{\theta}(\theta; data)\Big|_{\theta=\hat{\theta}_{MLE}} = \sum_{s=1}^{S} \frac{\partial \log f_s(y_s|X;\theta)}{\partial \theta}\Big|_{\theta=\hat{\theta}_{MLE}} = 0.$$

SOC:

$$\frac{\partial^2 \ell(\theta; data)}{\partial \theta \partial \theta'}\Big|_{\theta=\hat{\theta}_{MLE}} = \sum_{s=1}^{S} \ell_s^{\theta\theta'}(\theta; data)\Big|_{\theta=\hat{\theta}_{MLE}} = \sum_{s=1}^{S} \frac{\partial \log f_s(y_s|X;\theta)}{\partial \theta}\Big|_{\theta=\hat{\theta}_{MLE}}$$

is negative definite.

Note above that first order derivative of the log of the likelihood (also called the "score function") is a $p \times 1$ vector; and the second order derivative of the log of the likelihood is a $p \times p$ matrix.

Approximate SEV for MLE (whether linear/non-linear in y)

Focussing on the FOCs for a moment, we had under twice continuous differentiability of the likelihood function and the given model specification, that the ML estimator is defined by the following system of p equations in p unknowns:

FOC:

$$\sum_{s=1}^{S} \ell_s^{\theta}(\theta; data)\Big|_{\theta=\hat{\theta}_{MLE}} = 0.$$

Now, under $A5Gaussian$ for example, we can solve explicitly for $\hat{\theta}_{MLE}$. But what if we have a different $A5specific$, which although twice continuously differentiable, does not admit a closed-form expression for the ML estimator? In other words, what if the score function is non-linear in θ? (Ans: We find a linear approximation to the score function at θ^{true} and set that to zero!)

Consider the first order Taylor expansion of $\sum_{s=1}^{S} \ell_s^{\theta}(\theta; data)$ at θ^{true} given by

$$LHS = \sum_{s=1}^{S} \ell_s^{\theta}(\hat{\theta}_{MLE}; data) \approx \sum_{s=1}^{S} \ell_s^{\theta}(\theta^{true}; data) + \sum_{s=1}^{S} \ell_s^{\theta\theta'}(\theta^{true}; data)(\hat{\theta}_{MLE} - \theta^{true}) = RHS$$

The simple intuition is that since we cannot directly set LHS to zero and solve, we set a first order approximation of the LHS – i.e., the RHS – to zero and solve that instead.

Approximate SEV for MLE (whether linear/non-linear in y)

Continuing the analysis on the previous slide, we have that

$$\sum_{s=1}^{S} \ell_s^{\theta}(\theta^{true}; data) + \sum_{s=1}^{S} \ell_s^{\theta\theta'}(\theta^{true}; data)(\hat{\theta}_{MLE} - \theta^{true}) \approx 0,$$

so that by rearranging, we obtain

$$\left(\hat{\theta}_{MLE} - \theta^{true}\right) \approx -\left(\sum_{s=1}^{S} \ell_s^{\theta\theta'}(\theta^{true}; data)\right)^{-1} \sum_{s=1}^{S} \ell_s^{\theta}(\theta^{true}; data),$$

as the approximate SEV for ML estimator where the approximation is considered reasonable for sufficiently large sample size, S.

Asymptotic results for the OLS estimator in the ANLRM – scenario 1

Let us end this topic on asymptotics by summarising the asymptotic results available for the OLS estimator under the ANLRM. Notice that we are able to weaken our exogeneity assumption to "$\geq A3Rsru$" when considering results that are available only asymptotically as $S \to \infty$ (i.e., results that are not necessarily "exact").

Suppose we have ANLRM.$A4GM(iid)$ where $A1, A2linear, \geq A3Rsru, A4GM(iid)$. Then, we can/will prove (by LLNs, CLT and CMT2) that

$$\sqrt{S}(SEV(\hat{\beta}_{OLS})) = \sqrt{S}(\hat{\beta}_{OLS} - \beta^{true})|X \xrightarrow{d} N\left(0, \sigma_\varepsilon^2 \left(p \lim_{S\to\infty} \frac{X'X}{S}\right)^{-1}\right)$$

as $S \to \infty$.

Alternatively, for inferential purposes, the previous statement is taken as justification to say

$$\hat{\beta}_{OLS}|X \overset{approx}{\sim} N\left(\beta^{true}, \sigma_\varepsilon^2 (X'X)^{-1}\right)$$

for sufficiently large S.

Asymptotic results for the OLS estimator in the ANLRM – scenario 2

Suppose we have ANLRM.$A4\Omega$ where $A1, A2linear, \geq A3Rsru, A4\Omega$. Then, we can/will prove (by LLNs, CLT and CMT2) that

$$\sqrt{S}(SEV(\hat{\beta}_{OLS})) = \sqrt{S}(\hat{\beta}_{OLS} - \beta^{true})|X \xrightarrow{d} N\left(0, c^2 Q\right)$$

where

$$Q = \left(p \lim_{S\to\infty} \frac{X'X}{S}\right)^{-1} \left(p \lim_{S\to\infty} \frac{X'\Omega X}{S}\right) \left(p \lim_{S\to\infty} \frac{X'X}{S}\right)^{-1}$$

as $S \to \infty$.

Alternatively, for inferential purposes, the previous statement is taken as justification to say

$$\hat{\beta}_{OLS}|X \overset{approx}{\sim} N\left(\beta^{true}, c^2(X'X)^{-1}X'\Omega X(X'X)^{-1}\right)$$

for sufficiently large S.

Appreciating the distinction between different types of results

- Suppose we have NLRM.$A4GM(iid)$ where $A1, A2linear, \geq A3Rmi, A4GM(iid)$, and $A5Gaussian$ hold. Then, we can prove (due to preservation of multivariate Gaussianity) that
$$\hat{\beta}_{OLS}|X{\sim}N\left(\beta^{true}, \sigma_{\varepsilon}^{2}(X'X)^{-1}\right),$$
which is true for any S (even finite). This is an exact or finite-sample result true for any S.

- Suppose we have ANLRM.$A4GM(iid)$ where $A1, A2linear, \geq A3Rsru, A4GM(iid)$. Then, we can/will prove (by LLNs, CLT and CMT2) that
$$\sqrt{S}(SEV(\hat{\beta}_{OLS})) = \sqrt{S}(\hat{\beta}_{OLS} - \beta^{true})|X \xrightarrow{d} N\left(0, \sigma_{\varepsilon}^{2}\left(p\lim_{S\to\infty}\frac{X'X}{S}\right)^{-1}\right)$$
as $S \to \infty$. This is an asymptotically valid result true only as $S \to \infty$.

- Alternatively, for inferential purposes, the previous statement is taken as justification to say
$$\hat{\beta}_{OLS}|X \overset{approx}{\sim} N\left(\beta^{true}, \sigma_{\varepsilon}^{2}(X'X)^{-1}\right)$$
for sufficiently large S. This is an approximate result that is appropriate for large S.

12 *Asymptotic Theory — Technical Discussion [skim!]*

12.1 *Laws of Large Numbers (LLN)*

Weak LLN: Under suitable conditions, Sample moments converge in probability to their true population counterparts

Strong LLN: Under suitable conditions, Sample moments converge in MSE (alternatively almost surely) to their true population counterparts

12.2 *Central Limit Theorems (CLT)*

Under suitable conditions, Standardized Sample Mean will converge in distribution to a Gaussian random variable

12.3 *Continuous Mapping Theorems (CMT)*

12.3.1 *CMT1: Slutzsky*

Consider a (linear or nonlinear) function that is continuous in its two arguments, $g(\cdot,\cdot)$. Suppose that the first argument is a stochastic sequence Z_S in terms of S, with a finite probability limit $p\lim(Z_S) = Z^0_\infty$, and similarly that the second argument is a stochastic sequence W_S in terms of S, with a finite probability limit $p\lim(W_S) = W^0_\infty$. Then

$$g(Z_S, W_S) \quad \underset{\textbf{as } S\to\infty}{\overset{p}{\to}} \quad g(Z^0_\infty, W^0_\infty) = p\lim_{S\to\infty}\{g(Z_S, W_S)\}$$

NB1: The function $g(\cdot,\cdot)$ may be nonlinear or linear.
NB2: The two stochastic sequences may be statistically dependent or independent.

12.3.2 *CMT2: Cramér*

Consider a (linear or nonlinear) function that is continuous in its two arguments, $g(\cdot,\cdot)$. Suppose that the first argument is a stochastic sequence Z_S in terms of S, with a finite probability limit $p\lim(Z_S) = Z^0_\infty$, and similarly that the second argument is a stochastic sequence W_S in terms of S, which converges in distribution (as $S \to \infty$) to a random variable denoted by W^0. Then

$$g(Z_S, W_S) \quad \underset{\textbf{as } S \to \infty}{\overset{d}{\to}} \quad g(Z^0_\infty, W^0)$$

I.e., the asymptotic/limiting distribution of $g(Z_S, W_S)$ will be the same as that of the random variable $g(Z_S, W_S)$.

NB1: The function $g(\cdot, \cdot)$ may be nonlinear or linear.

NB2: The two stochastic sequences may be statistically dependent or independent.

NB3: Convergence is now in distribution, even though the first stochastic sequence converges in probability and the second in distribution.

12.4 *First Class of Asymptotic Results: Generically known as Laws of Large Numbers (LLNs):*

Definition 3 (Weak and Strong Laws of LLNs) ***Weak LLN (WLLN)****: Under certain conditions, the sample average (W_s) / sample first moment will* ***converge in probability*** *to the true population expectation underlying the true data-generating process.*

$W_s \rightarrow^p E(W_s)$ *as* $S \rightarrow \infty$

Strong LLN (SLLN)*: Under certain conditions, the sample average / sample first moment will converge in* ***Mean-Squared Error (MSE)*** *to the true population expectation underlying the true DGP.*

$W_s \rightarrow^{MSE,""AlmostSurely""} E(W_s)$ *as* $S \rightarrow \infty$

We are focusing on just two modes of convergence, in probability and MSE (ref. Billingsley, "*Convergence of statistical measures*").

If a WLLN applies to expression $\frac{1}{S}\sum_{s=1}^{S} a_s$, then $\frac{1}{S}\sum_{s=1}^{S} a_s \rightarrow^p E^{true}(a_s)$ as $S \rightarrow \infty$.

e.g. For OLS, $a_s \equiv x_s \epsilon_s^{true}$ which has:

$$E^{true}(a_s) = E(x_s \epsilon_s^{true}) = 0_{kx1} \text{ under A1+A2Linear+A3(Any)}, \therefore \frac{1}{S}\sum_{s=1}^{S} a_s \rightarrow^p E^{true}(a_s) = 0.$$

If a WLLN applies to expression $\frac{1}{S}\sum_{s=1}^{S} B_s$, then $\frac{1}{S}\sum_{s=1}^{S} B_s \rightarrow^p E(B_s)$ as $S \rightarrow \infty$, which by assumption, is a positive definite matrix by large sample implications of A1.

Definition 4 *Consistency*

If an estimator $\hat{\theta}_{method}$ *for true parameter* θ^{true}, *satisfies* $\hat{\theta}_{method} \rightarrow^p \theta^{true}$ *as* $S \rightarrow \infty$, *such an estimator is* ***weakly consistent*** *for* θ^{true}.

If an estimator $\hat{\theta}_{method}$ *for true parameter* θ^{true}, *satisfies* $\hat{\theta}_{method} \rightarrow^{MSE,AlmostSurely} \theta^{true}$ *as* $S \rightarrow \infty$, *such an estimator is* ***strongly consistent*** *for* θ^{true}.

Definition 5 *Convergence in MSE*

Define $q_S = \frac{1}{S}\sum_{s=1}^{S} a_{js}$ *="stochastic sequence in terms of S"; Something random that changes behavior when sample size S changes. In particular, it will have a pdf, true expectation, true variance, true interquartile range that changes with S.*

$q_s \rightarrow^{MSE} E(q_\infty)$ =*fixed number if and only if:*
$\lim_{S\rightarrow\infty} E(q_s) = E(q_\infty)$, *a fixed number.*
$\lim_{S\rightarrow\infty} Var(q_s) = 0$

Definition 6 *Convergence in Probability*
In contrast, convergence in probability $q_s \rightarrow^{p} E(q_\infty)$ =*fixed number if and only if:*
$\lim_{S\rightarrow\infty} \Pr \mid q_s \neq E(q_\infty) \mid= 0$

Explaining the difference in the two modes of convergence: From the definition of "strong" and "weak", it follows that if:

A stochastic sequence $\rightarrow^{MSE}_{S\rightarrow\infty}$to fixed number, this **implies** (same) stochastic sequence $\rightarrow^{p}_{S\rightarrow\infty}$fixed number. But why does this not go backwards? They are not the same because:

Convergence in probability holds more easily. Given an aim point in the sequence at the limit, with a small "blip"in the sequence stochastic sequence $\rightarrow^{p}$aim point since probability of sequence being different vanishes, but Var(stochastic sequence) blows up, which does not satisfy conditions for convergence in MSE.

12.5 *Second Class of Asymptotic Results: Central Limit Theorems (CLTs)*

Consider a stochastic sequence in S, $V_S = \frac{1}{\sqrt{S}}\sum_{s=1}^{S} V_s$, the normalized sample average.

$E(V_S) = 0 \forall s$ & $Var(V_S) = \sigma^2 < \infty \forall s$ and assume $V_S$drawn independently and identically. It follows that:

$E(V_S) = \frac{1}{\sqrt{S}} * S * 0 = 0$

$Var(V_S) = \left(\frac{1}{\sqrt{S}}\right)^2 * (S * \sigma^2 + 2Cov(\cdot)) = \frac{1}{S} S * \sigma^2 = \sigma^2$ ($Cov(\cdot) = 0$ since V_s is independently and identically drawn).

Summarizing: $V_S \sim ?(0, \sigma^2)$, where distribution will possibly change with S but will always have $E(V_s) = 0$, $Var(V_S) = \sigma^2$.

Definition 7 *The Central Limit Theorem states that under certain conditions, normalized sample average converges in distribution (as $S \to \infty$) to a Gaussian distribution:*

$V_S \sim ?(0, \sigma^2) \to^d_{S\to\infty} N(0, \sigma^2)$

Recall: Last part of normalized sampling error vector.

Define: A stochastic sequence in S, $V_S = \frac{1}{S}\sum_{s=1}^{S} a_s$.

For OLS, $a_s = x_s \epsilon_s^{true}$, $E(a_s) = 0_{kx1}$ under A1+A2Linear+A3(Any), $VCov(a_s) = E(x_s {\epsilon_s^{true}}^2 x_s') = \sigma_\epsilon^2 * E(x_s x_s')$ under A4GM(iid). Then, the appropriate CLT states that:

$$\frac{1}{\sqrt{S}}\sum_{s=1}^{S} a_s \to^d (0_{kx1}, \sigma_\epsilon^2 * E(x_s x_s')), as \frac{X'X}{S} \to^p_{S\to\infty} E(x_s x_s'), i.e. p\lim \frac{X'X}{S} = E(x_s x_s')$$

12.6 *Third Class of Asymptotic Results: Continuous Mapping Theorems (CMTs)*

Definition 8 *CMT1 (Slutsky): Consider some continuous (linear or non-linear function) $g(\cdot,\cdot)$ and consider two stochastic sequences in sample size S, Z_S and W_S that have valid probability limits:*

i.e. $Z_S \rightarrow^p_{S\rightarrow\infty} E(Z_\infty)$*, a non-stochastic number and* $W_S \rightarrow^p_{S\rightarrow\infty} E(W_\infty)$*, OR written as* $p\lim Z_S = E(Z_\infty)$ *and* $p\lim W_S = E(W_\infty)$ *.*

Note: Z_S may be independent of independent of W_S.

***Then** : CMT1 states that:*

$g(Z_S, W_s) \rightarrow^p_{S\rightarrow\infty} g(p\lim Z_s, p\lim W_s) = g(Z_\infty, W_\infty)$*, i.e.* $p\lim g(Z_S, W_s) = g(p\lim Z_s, p\lim W_s)$

CMT1 is useful because:

$\hat{\beta}_{method} - \beta^{true} = (\frac{1}{S}\sum_{s=1}^S B_s)^{-1}\frac{1}{S}\sum_{s=1}^S a_s$ where by a WLLN, $p\lim \frac{1}{S}\sum_{s=1}^S B_s = E(B_s)$, and $p\lim \frac{1}{S}\sum_{s=1}^S a_s = E(a_s) = 0$

*First equality if method is linear in y, approximate if non-linear in y.

$\therefore \hat{\beta}_{method} - \beta^{true} = (\frac{1}{S}\sum_{s=1}^S B_s)^{-1}\frac{1}{S}\sum_{s=1}^S a_s = g(\frac{1}{S}\sum_{s=1}^S B_s, \frac{1}{S}\sum_{s=1}^S a_s) = g(B_S, a_S)$ where $g(\cdot,\cdot)$ inverts the first argument and multiplies by the second. Expression can be defined by $g(\cdot,\cdot)$ because $g(\cdot,\cdot)$ is continuous in both arguments as each argument has a valid $p\lim$.

$\therefore g(\frac{1}{S}\sum_{s=1}^S B_s, \frac{1}{S}\sum_{s=1}^S a_s) \rightarrow^p_{S\rightarrow\infty} g(E(B_s), E(a_s)) = [E(B_s)]^{-1} * E(a_s)$.

Then:

$\hat{\beta}_{method} - \beta^{true} \rightarrow^p_{S\rightarrow\infty} [E(B_s)]^{-1} * E(a_s) = p\lim(\frac{X'X}{S})^{-1} * E(x_s\epsilon_s^{true}) = 0_{kx1}$.

Conclusion 9 *Under A1+A2Linear+A3(Any),* $\hat{\beta}_{OLS}$ *is a weakly consistent estimator for* β^{true}*, i.e.:*

$\hat{\beta}_{OLS} \rightarrow^p_{S\rightarrow\infty} \beta^{true}$

12.6.1 *Introducing CMT2 - Cramér:*

For any continuous (linear or non-linear) function with two arguments, $g(\cdot,\cdot)$ of two stochastic sequences, Z_s and V_s, which have the properties:

($\rightarrow^d$ &$\rightarrow^p$represent convergence in distribution / probability):

$Z_s \to^p Z_\infty^0$, i.e. $p\lim Z_s = Z_\infty^0$ (e.g. by appropriate law of large numbers)

$V_s \to^d V_\infty^0 \sim N(0, \sigma^2)$, e.g. if V_s is a normalized sample average (e.g. by appropriate CLT)

Note: Z_s and V_s may be dependent.

Then: $g(Z_s, V_s) \to^d g(p\lim Z_s, V_\infty^0)$ as $S \to \infty$, where $p\lim Z_s = Z_\infty^0$ and V_∞^0, the limiting random variable distribution. Here, V_s has a limiting distribution and becomes a Gaussian.

Result: For all methods discussed, under $A1, A2$,any $A3$, any $A4$, without $A5(Gaussian)$, whether approx. or exact, a (approximate) standard normal Gaussian linear regression model is obtained.

13 *Statistical Distributions related to NLRM: Gaussian, chi-squared, t-, and F- distributions [SKIM!]*

CASE I:	**NLRM with A4GM(iid)**
Result 1	Conditionally on X, $R\hat{\beta}_{OLS} \vert X \sim N\left(R\beta^{true}, R\left[\sigma^2_{\epsilon^{true}}(X'X)^{-1}\right]R'\right)$
Result 2	$\hat{s}^2_{OLS} \equiv \frac{RSS_{OLS}}{S-k}$is unbiased for $\sigma^2_{\epsilon^{true}}$ & conditionally on X, $\frac{(S-K)\hat{s}^2_{OLS}}{\sigma^2_{\epsilon^{true}}} \sim \chi^2(S-k)$
Result 3	Conditionally on X, $(R\hat{\beta}_{OLS} - R\beta^{true})'\left(R\left[\sigma^2_{\epsilon^{true}}(X'X)^{-1}\right]R'\right)^{-1}(R\hat{\beta}_{OLS} - R\beta^{true}) \sim \chi^2(r)$
Result 4A	Conditionally on X, the Gaussian r.v. $R\hat{\beta}_{OLS}$ of Result 1 and the χ^2 r.v. $\frac{(S-K)\hat{s}^2_{OLS}}{\sigma^2_{\epsilon^{true}}}$ of Result 2 are *independent*
Result 4B	Conditionally on X, theχ^2 r.v. $\frac{(S-K)\hat{s}^2_{OLS}}{\sigma^2_{\epsilon^{true}}}$of Result 2 and the χ^2 r.v. of Result 3 are *independent*
Result 5	Conditionally on X, $\tau \equiv \frac{\hat{\beta}^j_{OLS} - \beta^j_{true}}{\sqrt{\left[\hat{s}^2_{OLS}(X'X)^{-1}\right]_{jj}}} \sim t(S-k)$ because of Results 1,2,&4A
Result 6	Conditionally on X, $f \equiv (R\hat{\beta}_{OLS} - R\beta^{true})'\left(R\left[\hat{s}^2_{OLS}(X'X)^{-1}\right]R'\right)^{-1}(R\hat{\beta}_{OLS} - R\beta^{true})/r \sim F(r, S-k)$
Result 7	Suppose an estimator $\hat{\theta}_{method}$ for the $p \times 1$ unknown parameter vector θ^{true} obeys: $\hat{\theta}_{method} \sim N(\theta^{true}, V_{GM}(\hat{\theta}_{method}))$ for any sample size S; or $\hat{\theta}_{method} \approx N(\theta^{true}, V_{GM}(\hat{\theta}_{method}))$ for very large sample size S. Then for any continuous function $g(.) : R^p \rightarrow R^r$ with continuous first-derivative matrix $\left[\frac{\partial g(.)}{\partial \theta}\right]$ it follows that: $g(\hat{\theta}_{method}) \approx N\left(g(\theta^{true}), \left[\frac{\partial g(.)}{\partial \theta}\right] V_{GM}(\hat{\theta}_{method}) \left[\frac{\partial g(.)}{\partial \theta}\right]'\right)$ for very large S.

NOTES:

* In Result 2, $\hat{s}^2_{OLS}$ is both conditionally as well as unconditionally unbiased for $\sigma^2_{\epsilon^{true}}$ because we have proved $E\hat{s}^2_{OLS}|X = \sigma^2_{\epsilon^{true}}$, which means conditionally unbiased. By the Law of Iterated Expectations, we know this implies unconditionally unbiased also.

* Result 3 is the multivariate analogue of the result: "If $v \sim N(\mu, \sigma^2)$, then $z^2 \sim \chi^2(1)$ where $z \equiv \frac{v-\mu}{\sigma} \sim N(0,1)$"

* Result 4A follows because the $k \times S$ matrix of all cross-correlations between the r.v.s $R(\hat{\beta}_{OLS} - \beta^{true})$ and $\hat{\epsilon}_{ols} = M_X\epsilon^{true}$ equals:

$E\left[R(X'X)^{-1}X'\epsilon^{true}\epsilon^{true\prime}M_X|X\right] = R(X'X)^{-1}X'E\left[\epsilon^{true}\epsilon^{true\prime}|X\right]M_X$

$= \sigma^2_{\epsilon^{true}}R(X'X)^{-1}X'I_SM_X = 0.$

Since these two r.v.s are Gaussian distributed, uncorrelatedness is equivalent to independence. But squaring $\hat{\epsilon}_{ols}$ to get RSS_{ols} still means that independence will be preserved, so $R(\hat{\beta}_{OLS} - \beta^{true})$ will be fully independent of $\epsilon^{true\prime}M_X\epsilon^{true} = RSS_{ols}$ and $\hat{s}^2_{OLS}$.

* Result 4B is exactly analogous to 4A: having established the independence of $R(\hat{\beta}_{OLS}-\beta^{true})$ and $\hat{\epsilon}_{ols} = M_X\epsilon^{true}$, it follows that squaring each will give two r.v.s that are independent of each other. The square of the first is Result 3, while the square of the second is Result 2.

* Result 5 follows directly from the definition of the student-t distribution: "If r.v. $z \sim N(0,1)$ AND r.v. $q \sim \chi^2(df_q)$ AND z, q are independent, then $\tau \equiv \frac{z}{\sqrt{q/df_q}}$ $\sim$student-$t(df_q)$"

* Result 6 follows directly from the definition of the F distribution: "If r.v. $q \sim \chi^2(df_q)$ AND $q \sim \chi^2(df_q)$ AND w, q are independent, then $f \equiv \frac{w/df_w}{q/df_q} \sim F(df_w, df_q)$"

* Result 7 follows from the Taylor expansion that linearizes the $g(.)$ function: $g(\hat{\theta}_{method}) \approx g(\theta^{true}) + \frac{\partial g(\theta^{true})}{\partial \theta}(\hat{\theta}_{method} - \theta^{true}) + \cdots$, plus appropriate asymptotic arguments that hold as $S \to \infty$. This result is known as the "Delta method."

CASE II:	**NLRM with A4Ω**
Result 1	$R\hat{\beta}_{OLS} \mid X \sim N\left(R\beta^{true}, R\left[c^2(X'X)^{-1}X'\Omega X(X'X)^{-1}\right]R'\right)$
Result 2	$\hat{s}^2_{OLS} \equiv \frac{RSS_{OLS}}{S-k}$is NOT unbiased for $\sigma^2_{\epsilon^{true}}$ NOR for c^2 & $\frac{(S-K)\hat{s}^2_{OLS}}{\sigma^2_{\epsilon^{true}}}$ is NOT distributed $\chi^2(S-k)$ conditionally on X
Result 3	Conditionally on X, $(R\hat{\beta}_{OLS} - R\beta^{true})'\left(R\left[c^2(X'X)^{-1}X'\Omega X(X'X)^{-1}\right]R'\right)^{-1}(R\hat{\beta}_{OLS} - R\beta^{true}) \sim \chi^2(r)$
Result 4A	Conditionally on X, the Gaussian r.v. $R\hat{\beta}_{OLS}$ of Result 1 and the r.v. $\frac{(S-K)\hat{s}^2_{OLS}}{\sigma^2_{\epsilon^{true}}}$ of Result 2 are NOT *independent*
Result 4B	Conditionally on X, the r.v. $\frac{(S-K)\hat{s}^2_{OLS}}{\sigma^2_{\epsilon^{true}}}$of Result 2 and the χ^2 r.v. of Result 3 are NOT *independent*
Result 5	$\tau \equiv \frac{\hat{\beta}^j_{OLS}-\beta^j_{true}}{\sqrt{\left[\hat{s}^2_{OLS}(X'X)^{-1}\right]_{jj}}}$ is NOT distributed as $t(S-k)$ conditionally on X
Result 6	$f \equiv (R\hat{\beta}_{OLS} - R\beta^{true})'\left(R\left[c^2(X'X)^{-1}X'\Omega X(X'X)^{-1}\right]R'\right)^{-1}(R\hat{\beta}_{OLS} - R\beta^{true})/r$ is NOT distributed $F(.,.)$ given X.
Result 7	The Delta method for the A4Ω case is the same as in the A4GM case, except that $\hat{\theta}_{method}$ has VCov matrix V_Ω instead of V_{GM}.

NOTES:

* Result 2: The first part follows because now $ERSS_{ols}|X = trace(M_X E[\epsilon^{true}\epsilon^{true\prime}|X]) = c^2 trace(M_X\Omega)$. The second part follows because now ϵ^{true} is a Gaussian non-i.non-i.d. vector and hence squaring it does not give a χ^2 distribution.

* Result 3 follows because the correct normalization is performed so as to result in an i.i.d. standard Gaussian vector.

* Result 4A and 4B follow because now the r.v.s $R\hat{\beta}_{OLS}$ and $M_X\epsilon^{true}$ are no longer uncorrelated and hence they are statistically dependent.

* Result 5 follows because the r.v. $\frac{(S-K)\hat{s}^2_{OLS}}{\sigma^2_{\epsilon^{true}}}$ is (a) not distributed as χ^2 and (b) is not independent of the r.v. $R\hat{\beta}_{OLS}$.

* Result 6 follows because the r.v. $\frac{(S-K)\hat{s}^2_{OLS}}{\sigma^2_{\epsilon^{true}}}$ is (a) not distributed as χ^2 and (b) is not independent of the square of the r.v. $R\hat{\beta}_{OLS}$.

14 *Asymptotic/Large Sample Approximations of the Gaussian/ChiSquared/t-/F-Distributions (CMT2) [SKIM!]*

ANLRM with A4GM(iid)	
Result 1	Conditionally on X, $R\hat{\beta}_{OLS}\|X \approx N\left(R\beta^{true}, R\,[...]\,R'\right)$ for large S
Result 2	$\hat{s}^2_{OLS} \equiv \frac{RSS_{OLS}}{S-k} \overset{p}{\underset{\text{as } S \to \infty}{\rightarrow}} \sigma^2_{\epsilon^{true}}$ & $\frac{(S-K)\hat{s}^2_{OLS}}{\sigma^2_{\epsilon^{true}}} \sim \chi^2(\infty)$
Result 3	Conditionally on X, $(R\hat{\beta}_{OLS} - R\beta^{true})'\,(R\,[...]\,R')^{-1}\,(R\hat{\beta}_{OLS} - R\beta^{true}) \approx \chi^2(r)$ for large S
Result 4A	Conditionally on X, the Gaussian r.v. $R\hat{\beta}_{OLS}$ of Result 1 and the χ^2 r.v. $\frac{(S-K)\hat{s}^2_{OLS}}{\sigma^2_{\epsilon^{true}}}$ of Result 2 are *independent* Independence irrelevant for CMTs
Result 4B	Conditionally on X, theχ^2 r.v. $\frac{(S-K)\hat{s}^2_{OLS}}{\sigma^2_{\epsilon^{true}}}$of Result 2 and the χ^2 r.v. of Result 3 are *independent* Independence irrelevant for CMTs (e.g.,
Result 5	Conditionally on X, $\tau \equiv \frac{\hat{\beta}^j_{OLS} - \beta^j_{true}}{\sqrt{\left[\hat{s}^2_{OLS}(X'X)^{-1}\right]_{jj}}} \sim t(\infty) = N(0,1)$
Result 6	Conditionally on X, $f \equiv (R\hat{\beta}_{OLS} - R\beta^{true})'\,(R\,[\hat{s}^2_{OLS}(X'X)^{-1}]\,R')^{-1}\,(R\hat{\beta}_{OLS} - R\beta^{true})/r \sim F(r,\infty) = \chi^2(r)/r$
Result 7	Suppose an estimator $\hat{\theta}_{method}$ for the $p \times 1$ unknown parameter vector θ^{true} obeys: $\hat{\theta}_{method} \sim N(\theta^{true}, V_{GM}(\hat{\theta}_{method}))$ for any sample size S; or $\hat{\theta}_{method} \approx N(\theta^{true}, V_{GM}(\hat{\theta}_{method}))$ for very large sample size S. Then for any continuous function $g(.) : R^p \rightarrow R^r$ with continuous first-derivative matrix $\left[\frac{\partial g(.)}{\partial \theta}\right]$ it follows that: $g(\hat{\theta}_{method}) \approx N\left(g(\theta^{true}), \left[\frac{\partial g(.)}{\partial \theta}\right] V_{GM}(\hat{\theta}_{method}) \left[\frac{\partial g(.)}{\partial \theta}\right]'\right)$ for very large S. This result is already *asymp

14.1 *Establishing Result 7: The Delta Method [UNDERSTAND METHOD, SKIM DERIVATIONS!]*

Main Points: We will establish the asymptotic (as $S \to \infty$) distribution of $g(\hat{\beta}_{method})$, whether $\hat{\beta}_{method}$ ~Exactly $N(\cdot, \cdot)$ or Approximately $N(\cdot, \cdot)$ through the so-called Delta Method.

The Delta Method works through an asymptotic linearisation of $g(\cdot)$. Using a Taylor expansion around β^{true}:

$$g(\hat{\theta}) = g(\theta^{true}_{px1}) + \frac{\partial g(\theta^{true})}{\partial \theta}(\hat{\theta} - \theta^{true}) + Higher - orderterms.$$

Suppose that $\hat{\theta}_{method}$ is a CUAN estimator for θ^{true}_{px1}, i.e.:

$$\sqrt{S}(\hat{\theta} - \theta^{true}) \to^d N(0, VCov(\hat{\theta}_{method}))asS \to \infty \& \ \hat{\theta}_{method} \approx N(\theta^{true}, VCov(\hat{\theta}_{method}))forverylargeS.$$

In standard definition and notation:

$$\frac{\partial g(\cdot)}{\partial \theta} = \bigtriangledown_\theta g(\cdot) \equiv g_\theta(\cdot)$$

a r x p matrix of first derivative, $\bigtriangledown$represents the gradient function or"Delta" function.

Hence for very large S:

$$\sqrt{S}[g(\hat{\theta}) - g(\theta^{true})] \approx \bigtriangledown_\theta g(\theta^{true})\sqrt{S}(\hat{\theta}_{method} - \theta^{true})...Remainder\ (Leftout)$$

Since

$$\sqrt{S}(\hat{\theta}_{method} - \theta^{true}) \to^d N(0, VCov(\hat{\theta}_{method})),$$

$\therefore$By CMT2:

$$\sqrt{S}[g(\hat{\theta}) - g(\theta^{true})] \to^d \bigtriangledown_\theta g(\theta^{true}) * N(0, VCov(\hat{\theta}_{method}) = N(0_{rx1}, \bigtriangledown_\theta g(\theta^{true})VCov(\hat{\theta}_{method}) \bigtriangledown_\theta g(\theta^{true})')$$

Hence, the Delta Method gives the large sample approximation, since $\hat{\theta}_{method}$ is CUAN for θ^{true}:

$$g(\hat{\theta}_{method}) \approx N(g(\theta^{true}), \bigtriangledown_\theta g(\theta^{true})(\begin{pmatrix} VCov(\hat{\theta}_{method}) \\ S \end{pmatrix})[\bigtriangledown_\theta g(\theta^{true})]')$$

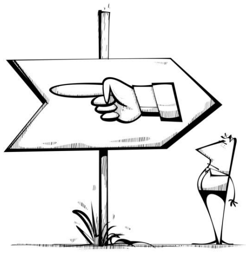
© Vassilis Hajivassiliou, LSE 2012–2023

Review Quiz for Topic 11

Question 1. What is convergence in probability? What is Slutsky's theorem (CMT1)? What is convergence in distribution? What is Cramér's theorem (CMT2)?

Question 2. Consider an n-th order polynomial $f(x)$ for some $n \geq 2$. Explain algebraically, graphically and intuitively what is meant by a first-order Taylor approximation of $f(x)$ at x_0.

Question 3. Explain why it would be incorrect to claim that:
“$\hat{\beta}_{OLS}|X \xrightarrow{d} N(\beta^{true}, \sigma^2_{\varepsilon}(X'X)^{-1})$ under $A1, A2linear, \geq A3Rsru, A4GM(iid)$ as $S \to \infty$.”



Signpost 11

Try to read the extract from Peter Kennedy's book, titled “Appendix C” (i.e., the pages that have been provided in the extended notes above BEFORE the main lecture slides). Then also try to read the supplementary technical notes from DrVH (i.e., the pages that have been provided in the extended notes above AFTER the main lecture slides).