Asymptotic Theory

L. Magee revised January 21, 2013

---

1 Convergence

1.1 Definitions

Let a_n to refer to a random variable that is a function of n random variables.

Convergence in Probability
The scalar a_n converges in probability to a constant α if, for any positive values of ϵ and δ, there is a sufficiently large n^* such that

$$\text{Prob}(|a_n - \alpha| > \epsilon) < \delta \text{ for all } n > n^*$$

When a_n converges in probability to α, then α is called the *probability limit*, or plim, of a_n.

Consistency
If a_n is an estimator of α, and plim $a_n = \alpha$, then a_n is a consistent estimator of α, or a little more briefly, a_n is consistent.

Convergence in Distribution
a_n converges in distribution to a random variable y ($a_n \to y$) if, as $n \to \infty$, $\text{Prob}(a_n \leq b) = \text{Prob}(y \leq b)$ for all b. Basically, the cumulative distribution function (and the probability density function) of a_n becomes the same as that of y as $n \to \infty$.

1.2 Functions of Variables that Converge

A nice feature of this convergence approach is that the properties of sums and products of random variables are much simpler to determine when they have converged. If a random variable has a probability limit, then as $n \to \infty$ it becomes nonrandom. If a random variable converges in probability, then in many standard cases it converges to a Normal or chi square distribution, which are well-known and have convenient properties.

Properties of Functions of Random Variables that Converge

(i) if $\text{plim}(x_n) = \theta_x$, then $\text{plim}(g(x_n)) = g(\theta_x)$, for any function $g(\cdot)$ that is continuous at θ_x. This is sometimes called *Slutsky's theorem.*
(For example, $\text{plim}(x_n^2) = \theta_x^2$, and $\text{plim}(1/x_n) = 1/\theta_x$ unless $\theta_x = 0$.

(ii) if x_n converges in distribution to some random variable x, i.e. $x_n \to x$, then, for any function $g(\cdot)$, $g(x_n) \to g(x)$. That is, the distribution of $g(x_n)$ converges to the distribution of $g(x)$. (This is like property (i), but for convergence in distribution instead of convergence in probability.)
(For example, if x_n converges in distribution to z, where z is a standard normal random variable (with a N[0,1] distribution) then x_n^2 converges in distribution to z^2, which has a chi square distribution with one degree of freedom.)

(iii) if $\text{plim}(x_n) = \theta_x$ and $\text{plim}(y_n) = \theta_y$, then $\text{plim}(x_n y_n) = \theta_x \theta_y$.
(Usually the distribution of products of random variables, like $x_n y_n$, are very complicated, but if x_n and y_n have plims, then at least it is easy to figure out the plim of $x_n y_n$.)

(iv) if $\text{plim}(x_n) = \theta_x$ and $y_n \to y$, then $x_n y_n \to \theta_x y$.
(This involves both a plim and a convergence in distribution. For example, if $\text{plim}(x_n) = \theta_x$ and $y_n \to z$, where $z \sim N[0,1]$, then $x_n y_n \to \theta_x z$, where $\theta_x z \sim N[0, \theta_x^2]$.) In the OLS example of Section 3, this sort of result is applied where "x_n" is a matrix, Slutsky's theorem (i) is applied to the matrix inverse function, and "y_n" converges to a normally distributed vector.

2 Theorems

There are many versions of these two theorems, with varying assumptions. In this handout these technical details are skipped to focus on the basic results.

2.1 Law of Large Numbers (LLN)

LLN is related to convergence in probability, and can be applied to a sum of random variables drawn from a distribution with a non-zero expected value. One version is:

If n random vectors $a_i, i = 1, \ldots, n$, are independently and identically distributed, with mean μ, then plim($n^{-1} \sum_i a_i = \mu$).

2.2 Central Limit Theorem (CLT)

The CLT originated with Laplace in 1810. It is the main theorem related to convergence in distribution. It can often be applied to a sum of random variables drawn from a distribution with zero expected value and finite variance. One version is:

If n random vectors $a_i, i = 1, \ldots, n$ are independently and identically distributed, with mean μ and variance Σ, then the distribution of $n^{1/2}(n^{-1}\sum_i a_i - \mu)$ converges to $N[0, \Sigma]$.

3 Asymptotic Properties of Ordinary Least Squares

3.1 Notation and Assumptions

Assume the true model is $y_i = x_i'\beta + u_i$, where x_i is a $k \times 1$ vector of observations on the RHS variables. x_i' is the i^{th} row of the usual $n \times k$ matrix X, and y_i is the i^{th} element of the usual $n \times 1$ vector y. So in this vector notation, a matrix product such as $X'X$ is written as $\sum x_i x_i'$.

Assume $Ex_iu_i = 0$ and define the $k \times k$ variance-covariance matrix of x_iu_i to be $\Sigma_{xu} \equiv \text{Var}(x_iu_i)$. Consider the OLS estimator of β, $b = (\sum x_i x_i')^{-1} \sum x_i y_i$. Unless indicated otherwise, the summations run over i from 1 to n, where n is the number of observations. Assume that the observations, the (x_i, y_i)'s, are random and independent across i, as in survey data where the randomness in both x_i and y_i derives from the random survey sampling. Substituting out y_i in the formula for b gives

$$b = \beta + (\sum x_i x_i')^{-1} \sum x_i u_i \tag{1}$$

3.2 Asymptotic Distribution of b

To find the asymptotic distribution of b, look at the two parts of the second RHS term of (1) separately. First, consider $\sum x_i x_i'$, a $k \times k$ matrix. Multiplying by n^{-1} gives a matrix $n^{-1}\sum x_i x_i'$ where each term in the matrix is a sample mean. Recalling that we assume that the vector x_i is random, let

$$\Sigma_{xx} \equiv E x_i x_i'$$

Applying LLN, then

$$\text{plim}(n^{-1} \sum x_i x_i') = \Sigma_{xx}$$

Assume that Σ_{xx} is invertible. This will be true if the elements in the x_i's are linearly independent and have finite variances. Applying a matrix version of Slutsky's theorem, then

$$\text{plim}(n^{-1} \sum x_i x_i')^{-1} = \Sigma_{xx}^{-1} \tag{2}$$

Second, apply the CLT to the last part of the second term in (1): $\sum x_i u_i$. To match this term to the way the CLT was given in subsection 2.2, note that

$$n^{-1/2} \sum x_i u_i = n^{1/2}(n^{-1} \sum_i x_i u_i - 0)$$

Since $Ex_i u_i = 0$, then the CLT implies

$$n^{1/2}(n^{-1} \sum_i x_i u_i - 0) \rightarrow N[0, \Sigma_{xu}] \tag{3}$$

where Σ_{xu} is the population variance-covariance matrix of the $x_i u_i$'s. Since $Ex_i u_i = 0$, then $\Sigma_{xu} = E(x_i u_i)(x_i u_i)' = Eu_i^2 x_1 x_i'$. Next, put b from (1) in a form in which (2) and (3) can be applied.

$$\begin{aligned} b - \beta &= (\sum x_i x_i')^{-1} \sum x_i u_i \\ &= (n^{-1} \sum x_i x_i')^{-1} (n^{-1} \sum x_i u_i) \\ n^{1/2}(b - \beta) &= (n^{-1} \sum x_i x_i')^{-1} (n^{1/2}(n^{-1} \sum x_i u_i - 0)) \end{aligned}$$

and apply (2) and (3) to the two terms in parentheses, respectively,

$$\begin{aligned} n^{1/2}(b - \beta) &\rightarrow (\Sigma_{xx}^{-1}) \times (\text{a } N[0, \Sigma_{xu}] \text{ random variable vector}) \\ &\rightarrow N[0, \Sigma_{xx}^{-1} \Sigma_{xu} \Sigma_{xx}^{-1}] \end{aligned} \tag{4}$$

3.3 Implementing the Asymptotic Result

Let the symbol "≈" represent "approximately distributed as". Then it follows from (4) that

$$\begin{aligned} n^{1/2}(b-\beta) &\approx N[0, \Sigma_{xx}^{-1}\Sigma_{xu}\Sigma_{xx}^{-1}] \\ b-\beta &\approx N[0, n^{-1}\Sigma_{xx}^{-1}\Sigma_{xu}\Sigma_{xx}^{-1}] \\ b &\approx N[\beta, n^{-1}\Sigma_{xx}^{-1}\Sigma_{xu}\Sigma_{xx}^{-1}] \end{aligned} \tag{5}$$

To use this result for inference, an estimator for the variance-covariance matrix $n^{-1}\Sigma_{xx}^{-1}\Sigma_{xu}\Sigma_{xx}^{-1}$ is required. Replace Σ_{xx} and Σ_{xu} by consistent estimators as follows. Since $\Sigma_{xx} = Ex_ix_i'$ is the population mean of the x_ix_i' matrices, a natural estimator of Σ_{xx} is the sample mean of the observed x_ix_i' matrices,

$$\hat{\Sigma}_{xx} = n^{-1}\sum_i x_ix_i'$$

Applying Slutsky's theorem to the matrix inverse function, both sides can be inverted to give

$$\hat{\Sigma}_{xx}^{-1} = (n^{-1}\sum_i x_ix_i')^{-1}$$

The other matrix we need to estimate, Σ_{xu}, is the population variance-covariance matrix of the x_iu_i vectors. Since $E(x_iu_i) = 0$, then

$$\begin{aligned} \Sigma_{xu} &= \mathrm{Var}(x_iu_i) \\ &= E(x_iu_i - E(x_iu_i))(x_iu_i - E(x_iu_i))' \\ &= E(x_iu_i - 0)(x_iu_i - 0)' \\ &= E(x_iu_i)(x_iu_i)' \\ &= E(x_iu_iu_i'x_i') \\ &= E(u_i^2x_ix_i') \end{aligned}$$

where the last line follows from the fact that u_i is a scalar. We can interpret Σ_{xu} then as the population mean of the $u_i^2x_ix_i'$'s. If we knew the sample u_i's, then it would be natural to estimate Σ_{xu} by the sample mean of the $u_i^2x_ix_i'$'s. While we do not know u_i, we can replace it with the OLS residual $e_i = y_i - x_i'b$. Since $\mathrm{plim}(b) = \beta$, then $\mathrm{plim}(e_i) = y_i - x_i'\beta = u_i$, making this replacement valid in an asymptotic framework, enabling us to estimate Σ_{xu} by the sample mean of the $e_i^2 x_i x_i'$'s:

$$\hat{\Sigma}_{xu} = n^{-1} \sum_i e_i^2 x_i x_i'$$

Substituting these two matrix estimators into the approximate variance of b given in (5) produces what is known as White's "HCCME" (heteroskedasticity-consistent covariance matrix estimator):

$$\begin{aligned} \widehat{\text{Var}}(b) &= n^{-1}\hat{\Sigma}_{xx}^{-1}\hat{\Sigma}_{xu}^{-1}\hat{\Sigma}_{xx}^{-1} \\ &= n^{-1}(n^{-1}\sum_i x_i x_i')^{-1}(n^{-1}\sum_i e_i^2 x_i x_i')(n^{-1}\sum_i x_i x_i')^{-1} \\ &= (\sum_i x_i x_i')^{-1}(\sum_i e_i^2 x_i x_i')(\sum_i x_i x_i')^{-1} \qquad (6) \end{aligned}$$

Summarizing, under the assumptions given in subsection 3.1, the distribution of b is approximately

$$b \approx N[\beta, \widehat{\text{Var}}(b)]$$

The assumptions in 3.1 are weaker than the classical linear regression model in several ways. First, they allow for heteroskedastic errors. Although $\text{Var}(x_i u_i) = \Sigma_{xx}$ may appear to be a no-heteroskedasticity assumption because Σ_{xx} is constant across i, it does not rule out variation in the variance of u_i conditional on x_i. That is, we can still have $\text{Var}(u_i|x_i = x_A) \neq \text{Var}(u_i|x_i = x_B)$ when $x_A \neq x_B$.

Second, these assumptions do not require that $E(y_i|x_i) = x_i'\beta$ for all i. In other words, we do not have to assume knowledge of the functional form of the regression function, the population mean of y_i given x_i. We have assumed that $E(x_i u_i) = 0$, which implies $E(x_i(y_i - x_i\beta) = 0$, or $E(x_i y_i) = Ex_i x_i'\beta$. This allows for $E(y_i|x_i) > x_i'\beta$ for some x_i's, and $E(y_i|x_i) < x_i'\beta$ for other x_i's, as long as the weighted expected values of the two sides of these inequalities, $E(x_i y_i)$ and $Ex_i x_i'\beta$, are equal. This means that we can use OLS and estimate its variance in this framework without even knowing the functional form of the true regression function $E(y_i|x_i)$, although we must then acknowledge that we are only estimating the parameters (β) of an approximation to the true regression function (the approximation being the linear model $x_i'\beta$).

Third, we do not need to assume that the error terms, u_i, are normally distributed.