

Lecture Note 1: Brief Review of Asymptotics

(Drawn from van der Vaart, Ch. 2., Wooldridge, Ch.3, and Severini Ch. 11, 12.1-12.5, 13.2.)

Random Variables: we will use the term *random variable* to refer to random vectors in \mathbb{R}^k . The *distribution function* of a random variable X is the function

$$F(x) = P(X \leq x).$$

Convergence Concepts

Convergence of Deterministic Sequences: a sequence of nonrandom numbers $X_i : i = 1, 2, \dots$ converges to a limit a if, for any $\epsilon > 0$, there exists n_ϵ such if $n > n_\epsilon$ then $|X_n - a| < \epsilon$. We write $X_n \rightarrow a$ as $n \rightarrow \infty$.

Convergence in Probability: a sequence of random variables $\{X_i\}$ converges in probability to X if, for all $\epsilon > 0$,

$$P(\|X_i - X\| > \epsilon) \rightarrow 0.$$

This is denoted by $X_i \xrightarrow{p} X$, and we also write that $X = \text{plim } X_i$.

Usually, the limit X will be a constant, but random limits are allowed in the definition. When X_i is a vector and its probability limit is a constant, the definition above is equivalent to convergence in probability of each of the elements of X_i .

Convergence in Distribution: a sequence of random variables $\{X_i\}$ is said to converge in distribution to a random variable X if

$$P(X_i \leq x) \rightarrow P(X \leq x)$$

at every point x at which the limit distribution function $P(X \leq x)$ is continuous. This will be denoted by $X_i \xrightarrow{d} X$, or sometimes $X_i \rightsquigarrow X$.

Basic Asymptotic Results:

Weak Law of Large Numbers (WLLN): Let X_1, X_2, \dots be a sequence of i.i.d. random variables such that $E(\|X_1\|) < \infty$. Then

$$\bar{X}_n \equiv \frac{1}{n} \sum_i X_i \xrightarrow{p} E(X_1).$$

Multivariate Central Limit Theorem (CLT): Let X_1, X_2, \dots be i.i.d. random vectors in \mathbb{R}^k with mean $\mu = EX_1$ and covariance matrix $\Sigma = E(X_1 - \mu)(X_1 - \mu)'$. Then

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n (X_i - \mu) = \sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} N(0, \Sigma).$$

Continuous Mapping Theorem (CMT): Let $g(\cdot)$ be a function from \mathbb{R}^k to \mathbb{R}^m , and suppose g is continuous at every point in a set C s.t. $P(X \in C) = 1$. Then

- (i) If $X_n \xrightarrow{p} X$, then $g(X_n) \xrightarrow{p} g(X)$;
- (ii) If $X_n \xrightarrow{d} X$, then $g(X_n) \xrightarrow{d} g(X)$.

Note that matrix addition and multiplication are continuous functions.

There are various useful facts about convergence in probability and convergence in distribution:

Result:

- (i) Convergence in probability implies convergence in distribution:

$$X_i \xrightarrow{p} X \Rightarrow X_i \xrightarrow{d} X.$$

- (ii) Convergence in distribution to a constant, implies convergence in probability:

$$X_i \xrightarrow{d} a \text{ (a constant)} \Rightarrow X_i \xrightarrow{p} a.$$

- (iii) If $X_n \xrightarrow{d} X$ and $\|X_n - Y_n\| \xrightarrow{p} 0$, then $Y_n \xrightarrow{d} X$.
- (iv) If $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{p} a$ where a is a constant, then the vector $(X_n, Y_n) \xrightarrow{d} (X, a)$.
- (v) If $X_n \xrightarrow{p} X$ and $Y_n \xrightarrow{p} Y$, then $(X_n, Y_n) \xrightarrow{p} (X, Y)$. (This is not true for convergence in distribution.)

A useful corollary of the previous result:

Slutsky's Lemma: Let X_n, X , and Y_n be random vectors or matrices. If $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{d} c$, where c is a constant, then

- (i) $X_n + Y_n \xrightarrow{d} X + c$;
- (ii) $Y_n X_n \xrightarrow{d} cX$;

(iii) $Y_n^{-1}X_n \xrightarrow{d} c^{-1}X$, provided c is invertible.

Delta Method: Let X_n be a sequence of d -dimensional random vectors such that

$$\sqrt{n}(X_n - \mu) \xrightarrow{d} N(0, \Sigma),$$

where Σ is positive definite and finite. Let g denote a continuously differentiable function from \mathbb{R}^d into \mathbb{R}^k , and let $G(x) = \partial g / \partial x$ denote the $k \times d$ matrix of partial derivatives. Then

$$\sqrt{n}(g(X_n) - g(\mu)) \xrightarrow{d} N(0, G(\mu)\Sigma G(\mu)').$$

o and O notation: Let a_1, a_2, \dots and b_1, b_2, \dots denote sequences of real numbers, where $b_n > 0$ for all n . The notation $a_n = O(b_n)$ means that there exists a constant M such that $|a_n| \leq Mb_n$ for all n .

The notation $a_n = o(b_n)$ means that $\lim_{n \rightarrow \infty} a_n/b_n = 0$.

o_p and O_p notation:

Let X_n denote a sequence of scalar random variables. Suppose that for any $\epsilon > 0$ there exists a constant M such that

$$P(|X_n| \geq M) \leq \epsilon$$

for all n . Then we write $X_n = O_p(1)$. Note that if X_n converges in distribution to some random variable, then it is $O_p(1)$. Another sufficient condition is that for some increasing function g ,

$$\sup_n E[g(|X_n|)] < \infty.$$

If $X_n \xrightarrow{p} 0$, we write $X_n = o_p(1)$.

Now, let X_n and Y_n denote sequences of random variables, where $Y_n > 0$ for all n . We write $X_n = O_p(Y_n)$ if

$$\frac{X_n}{Y_n} = O_p(1).$$

We write $X_n = o_p(Y_n)$ if

$$\frac{X_n}{Y_n} = o_p(1).$$

There are a number of simple rules for working with o_p and O_p random variables, including:

- If $X_n = o_p(1)$, then $X_n = O_p(1)$.
- If $X_n = O_p(1)$ and $W_n = O_p(1)$, then $X_n + W_n = O_p(1)$ and $X_n W_n = O_p(1)$. We write this as $O_p(1) + O_p(1) = O_p(1)$, and $O_p(1) \cdot O_p(1) = O_p(1)$.
- $O_p(1) \cdot o_p(1) = o_p(1)$ and $O_p(1) + o_p(1) = O_p(1)$.
- If $X_n = O_p(Y_n)$ and $W_n = O_p(Z_n)$, then $W_n X_n = O_p(Y_n Z_n)$, and $W_n + X_n = O_p(\max(Z_n, Y_n))$.
- $(1 + o_p(1))^{-1} = O_p(1)$.
- $o_p(O_p(1)) = o_p(1)$.

Estimators and Test Statistics

Let $\hat{\theta}_n$ be a sequence of estimators of $\theta \in \Theta \subset \mathbb{R}^k$, where n indexes the sample size. If $\hat{\theta}_n \xrightarrow{p} \theta$, for any value of θ , we say that $\hat{\theta}$ is a *consistent estimator* of θ . Note that in practice, we often drop the n subscript from $\hat{\theta}$.

Suppose further that

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} N(0, V),$$

where V is a $k \times k$ positive semidefinite matrix. Then we say that $\hat{\theta}_n$ is *\sqrt{n} -asymptotically normally distributed* and V is the *asymptotic variance* of $\sqrt{n}(\hat{\theta}_n - \theta)$.

Loosely speaking, we take this to mean that the approximate distribution $\hat{\theta}_n$ is $N(\theta, V/n)$. So the square root of the (j, j) th element of V/n would be the standard deviation of $\hat{\theta}_{nj}$ (the j th element of $\hat{\theta}_n$).

In practice, we don't know V exactly, but we can usually estimate it consistently:

$$\hat{V}_n \xrightarrow{p} V.$$

Letting \hat{V}_{njj} denote the j, j element of \hat{V}_n , we estimate the standard deviation of $\hat{\theta}_{nj}$ by $\sqrt{\hat{V}_{njj}/n}$. We call this the *standard error* of $\hat{\theta}_{nj}$. Usually, we report parameter estimates along with their standard errors. With these two pieces, we can easily form confidence intervals and simple t statistics.

Some definitions related to test statistics:

The *asymptotic size* of a testing procedure is the limiting probability of rejecting the null hypothesis

(H_0) when it is true. We can write this as:

$$\lim_{n \rightarrow \infty} P_n(\text{reject } H_0 | H_0 \text{ is true}),$$

where P_n means the conditional probability for sample size n .

A test is said to be *consistent* against the alternative H_1 if the null hypothesis is rejected with probability approaching one if H_1 is true:

$$\lim_{n \rightarrow \infty} P_n(\text{reject } H_0 | H_1) = 1.$$

Suppose we want to test the restriction that

$$H_0 : R\theta = r,$$

where R is a $q \times k$ matrix with $q \leq k$ and rank q , and r is a $q \times 1$ vector. Let us derive the standard Wald test. Assume that we have an estimator $\hat{\theta}$ that is consistent and \sqrt{n} -asymptotically normal, with

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N(0, V).$$

By the continuous mapping theorem and the properties of the multivariate normal distribution,

$$\sqrt{n}R(\hat{\theta} - \theta) \xrightarrow{d} N(0, RV R'),$$

and

$$[\sqrt{n}R(\hat{\theta} - \theta)]' [RV R']^{-1} [\sqrt{n}R(\hat{\theta} - \theta)] \xrightarrow{d} \chi_q^2.$$

In addition, if $\hat{V} \xrightarrow{p} V$, then

$$\begin{aligned} [\sqrt{n}R(\hat{\theta} - \theta)]' [R\hat{V}R']^{-1} [\sqrt{n}R(\hat{\theta} - \theta)] &= (R\hat{\theta} - R\theta)' [R(\hat{V}/n)R']^{-1} (R\hat{\theta} - R\theta) \\ &\xrightarrow{d} \chi_q^2. \end{aligned}$$

So, to test the null hypothesis that $R\theta = r$, we define the Wald statistic as

$$W_n = (R\hat{\theta} - r)' [R(\hat{V}/n)R']^{-1} (R\hat{\theta} - r)$$

and compare its value to critical values for the χ_q^2 distribution.

We can extend this approach to test hypotheses about nonlinear functions of θ . Suppose that $c : \Theta \rightarrow \mathbb{R}^q$ is a continuously differentiable function with $q \leq k$, and we are interested in testing the hypothesis that $c(\theta) = 0$. Assume that θ is in the interior of Θ . Then, by the delta method,

$$\sqrt{n}(c(\hat{\theta}) - c(\theta)) \xrightarrow{d} N(0, C(\theta)VC(\theta)'),$$

where $C(\theta) = \partial c(\theta)/\partial \theta$ is the $q \times p$ Jacobian of c . Then

$$[\sqrt{n}(c(\hat{\theta}) - c(\theta))]'[C(\theta)VC(\theta)']^{-1}[\sqrt{n}(c(\hat{\theta}) - c(\theta))] \xrightarrow{d} \chi_q^2.$$

Let $\hat{C} = C(\hat{\theta})$. By assumption, C is continuous, so by the continuous mapping theorem $\hat{C} \xrightarrow{p} C(\theta)$. If we also have a consistent estimator \hat{V} of V , then

$$[\sqrt{n}(c(\hat{\theta}) - c(\theta))]')[\hat{C}\hat{V}\hat{C}']^{-1}[\sqrt{n}(c(\hat{\theta}) - c(\theta))] \xrightarrow{d} \chi_q^2.$$

Under the null hypothesis, $c(\theta) = 0$, so the appropriate Wald statistic would be:

$$\begin{aligned} W_n &= [\sqrt{nc}(\hat{\theta})]'[\hat{C}\hat{V}\hat{C}']^{-1}[\sqrt{nc}(\hat{\theta})] \\ &= c(\hat{\theta})'[\hat{C}(\hat{V}/n)\hat{C}']^{-1}c(\hat{\theta}), \end{aligned}$$

which is asymptotically distributed as χ_q^2 under H_0 .

Example: Best Linear Predictor

Suppose that $Z_i = (Y_i, X_i)$ is an IID random vector, with Y_i scalar and X_i a k -vector. Suppose that Z_i has bounded second moments, $E[X_i X_i']$ is nonsingular, and that the elements of X_i are linearly independent with positive probability. Consider the following problem:

$$\min_{\gamma \in \mathbb{R}^k} E[(Y_i - X_i' \gamma)^2].$$

If the conditional mean $E[Y_i|X_i]$ is linear, that is $E[Y_i|X_i] = X_i' \beta$, then the solution is $\gamma = \beta$.

Even if the conditional mean is not linear, then γ solving the best linear predictor problem is still defined, and can be shown¹ to be

$$\gamma = (E[X_i X_i'])^{-1} E[X_i Y_i].$$

Consider the least squares estimator

$$\hat{\beta} = \left(\sum_{i=1}^n X_i X_i' \right)^{-1} \sum_{i=1}^n X_i Y_i.$$

Exercise: Without assuming that $E[Y_i|X_i]$ is linear in X_i , show that $\hat{\beta} \xrightarrow{p} \gamma$ and derive its limiting distribution, making any additional assumptions necessary.

¹For example, see Ruud, P., *An Introduction to Classical Econometric Theory*, Lemma 7.4.