The Uniform Weak Law of Large Numbers and the Consistency of M-Estimators of Cross-Section and Time Series Models

Herman J. Bierens

Pennsylvania State University

September 16, 2005

1. *The uniform weak law of large numbers*

In econometrics we often have to deal with sample means of random functions. A random function is a function that is a random variable for each fixed value of its argument. In cross-section econometrics random functions usually take the form of a function $g(Z,\theta)$ of a random vector Z and a non-random vector θ. For example, consider a Logit model:

$$P[Y_j = y|X_j] = \frac{y + (1 - y)\exp(-\alpha - \beta^T X_j)}{1 + \exp(-\alpha - \beta^T X_j)},\ y = 0,1,$$

where $Y_j \in \{0,1\}$ is the dependent variable and $X_j \in \mathbb{R}^k$ is a vector of explanatory variables. Denoting $Z_j = (Y_j, X_j^T)^T$, and given a random sample $\{Z_1, Z_2, ..., Z_n\}$, the log-likelihood function involved takes the form $\sum_{j=1}^n g(Z_j, \theta)$, where

$$\begin{aligned} g(Z_j,\theta) &= \ln\left(Y_j + (1 - Y_j)\exp(-\alpha - \beta^T X_j)\right) - \ln\left(1 + \exp(-\alpha - \beta^T X_j)\right) \\ &= Y_j(\alpha + \beta^T X_j) - \ln\left(1 + \exp(\alpha + \beta^T X_j)\right),\ \text{where } \theta = (\alpha, \beta^T)^T. \end{aligned} \tag{1}$$

For such functions we can extend the weak law of large numbers for i.i.d. random variables to a Uniform Weak Law of Large Numbers (UWLLN):

Theorem 1: *Let Z_j, $j = 1,..,n$, be a random sample from a k-variate distribution. Let $g(z,\theta)$ be a Borel measurable function on $\mathbf{Z} \times \Theta$, where $\mathbf{Z} \subset \mathbb{R}^k$ is a Borel set such that $P[Z_t \in \mathbf{Z}] = 1$, and Θ is a compact subset of $\mathbb{R}^m$, such that for each $z \in \mathbf{Z}$, $g(z,\theta)$ is a continuous function on Θ. Furthermore, let*

$$E[\sup_{\theta\in\Theta}|g(Z_j,\theta)|] < \infty. \qquad (2)$$

Then $\text{plim}_{n\to\infty}\sup_{\theta\in\Theta}|(1/n)\sum_{j=1}^{n}g(Z_j,\theta) - E[g(Z_1,\theta)]| = 0$.

Note that subsets of Euclidean spaces are compact if and only if they are closed and bounded. See, for example, Bierens (2004), Appendix II, Theorem II.2.

The original proof of the stronger result

$$\sup_{\theta\in\Theta}|(1/n)\sum_{j=1}^{n}g(Z_j,\theta) - E[g(Z_1,\theta)]| \to 0 \text{ a.s.},$$

was given in the seminal paper of Jennrich (1969). This proof is explained in detail in Bierens (2004, Appendix to Chapter 6).

The condition that the random vectors Z_j are i.i.d. can be relaxed, because the result in Theorem 1 also holds for strictly stationary time series processes with a vanishing memory:

Definition 1: *A* (*vector*) *time series process* $X_t \in \mathbb{R}^k$ *is strictly stationary if for arbitrary integers* $m_1 < m_2 < ....< m_n$ *the joint distribution of* $\left(X_{t-m_1}^T,.....,X_{t-m_n}^T\right)^T$ *does not depend on the time index t.*

Definition 2: *A* (*vector*) *time series process* $X_t \in \mathbb{R}^k$ *has a vanishing memory if all the sets in the remote* σ*-algebra* $\mathscr{F}_{-\infty} = \bigcap_t \sigma\left(\{X_{t-j}\}_{j=0}^{\infty}\right)$ *have either probability zero or one.*

Note that if the X_t's are independent then by Kolmogorov's zero-one law the time series X_t has a vanishing memory.

It has been shown in Bierens (2004, Theorem 7.4) that

Theorem 2: *If* $X_t \in \mathbb{R}^k$ *is a strictly stationary time series process with vanishing memory, and* $E[\|X_t\|] < \infty$, *then* $\text{plim}_{n\to\infty}(1/n)\sum_{t=1}^{n}X_t = E[X_1]$.

I will use this result to prove the following more general version of Theorem 1. To be able to generalize the UWLLN to the time series case where the random functions involved depend on

the entire past of the time series rather than on a finite dimensional vector of variables, I will reformulate and prove Theorem 1 under slightly different moment conditions.

Theorem 3: *Let* $Z_t \in \mathbb{R}^k$ *be a strictly stationary vector time series process with a vanishing memory,*[1] *defined on a common probability space* $\{\Omega, \mathscr{F}, P\}$. *Let* $g(z,\theta)$ *be a Borel measurable real function on* $\mathbf{Z} \times \Theta_0$, *where* $\mathbf{Z} \subset \mathbb{R}^k$ *is a Borel set such that* $P[Z_t \in \mathbf{Z}] = 1$, *and* Θ_0 *is an open subset of* $\mathbb{R}^m$, *such that for each* $z \in \mathbf{Z}$, $g(z,\theta)$ *is a continuous function on* Θ_0. *Furthermore, let* Θ *be a compact subset of* Θ_0. *Finally, assume that for each* $\theta_* \in \Theta$ *there exists an arbitrary small* $\delta > 0$, *possibly depending on* θ_*, *such that*

$$E\left[\sup_{\|\theta-\theta_*\| \le \delta} g(Z_1,\theta)\right] < \infty, \quad E\left[\inf_{\|\theta-\theta_*\| \le \delta} g(Z_1,\theta)\right] > -\infty. \tag{3}$$

Then $\operatorname{plim}_{n\to\infty}\sup_{\theta\in\Theta}|(1/n)\sum_{j=1}^{n} g(Z_j,\theta) - E[g(Z_1,\theta)]| = 0$.

Proof: Observe from condition (3) that for each $\theta \in \Theta$, $E[g(Z_1,\theta)]$ is well-defined. Actually, due to the compactness of Θ, (3) implies (2) [*Exercise*: Why?], so that the latter is a weaker condition than (3). Moreover, it follows from condition (3), the continuity of $g(z,\theta)$ in θ, and the dominated convergence theorem, that

$$\lim_{\delta\downarrow 0} E\left[\sup_{\|\theta-\theta_*\| \le \delta} g(Z_1,\theta) - \inf_{\|\theta-\theta_*\| \le \delta} g(Z_1,\theta)\right] = 0, \tag{4}$$

pointwise in $\theta_* \in \Theta$. Therefore, for an arbitrary $\varepsilon > 0$ and each $\theta_* \in \Theta$ we can choose a positive number $\delta(\theta_*,\varepsilon)$ such that, with

$$N(\theta_*|\varepsilon) = \{\theta \in \Theta_0: \|\theta - \theta_*\| < \delta(\theta_*,\varepsilon)\}, \tag{5}$$

we have

$$0 \le E\left[\sup_{\theta \in N(\theta_*|\varepsilon)} g(Z_1,\theta) - \inf_{\theta \in N(\theta_*|\varepsilon)} g(Z_1,\theta)\right] < \varepsilon. \tag{6}$$

Next, observe that the sets (5) are open, so that $\bigcup_{\theta_*\in\Theta} N(\theta_*|\varepsilon)$ is an open covering of Θ. Then by the compactness of Θ there exists a finite sub-covering of Θ:

---

[1] Which includes the case that the Z_t's are i.i.d.

$$\Theta \subset \bigcup_{i=1}^{K} N(\theta_i|\varepsilon), \tag{7}$$

where K and the vectors $\theta_i \in \Theta$ depend on ε.

Using the easy inequality $\sup_x|f(x)| \leq |\sup_x f(x)| + |\inf_x f(x)|$, it is not hard to verify that for each $\theta_i \in \Theta$,

$$\begin{aligned}
&\sup_{\theta\in N(\theta_i|\varepsilon)}|(1/n)\textstyle\sum_{t=1}^{n} g(Z_t,\theta) - E[g(Z_1,\theta)|\\
&\leq 2|(1/n)\textstyle\sum_{t=1}^{n}\sup_{\theta\in N(\theta_i|\varepsilon)} g(Z_t,\theta) - E[\sup_{\theta\in N(\theta_i|\varepsilon)} g(Z_1,\theta)]|\\
&+ 2|(1/n)\textstyle\sum_{t=1}^{n}\inf_{\theta\in N(\theta_i|\varepsilon)} g(Z_t,\theta) - E[\inf_{\theta\in N(\theta_i|\varepsilon)} g(Z_1,\theta)]|\\
&+ 2\Big(E[\sup_{\theta\in N(\theta_i|\varepsilon)} g(Z_1,\theta)] - E[\inf_{\theta\in N(\theta_i|\varepsilon)} g(Z_1,\theta)]\Big).
\end{aligned} \tag{8}$$

It follows from Theorem 2 that the first two terms at the right-hand side of (8) converge in probability to zero, and from (6) that the last term is less than $2.\varepsilon$. Hence,

$$\begin{aligned}
&\sup_{\theta\in\Theta}|(1/n)\textstyle\sum_{t=1}^{n} g(Z_t,\theta) - E[g(Z_1,\theta)|\\
&\leq \max_{1\leq i\leq K}\sup_{\theta\in N(\theta_i|\varepsilon)}|(1/n)\textstyle\sum_{t=1}^{n} g(Z_t,\theta) - E[g(Z_1,\theta)|\\
&\leq R_n(\varepsilon) + 2.\varepsilon, \text{ where } \operatorname{plim}_{n\to\infty} R_n(\varepsilon) = 0.
\end{aligned} \tag{9}$$

Theorem 3 follows now straightforwardly from (9). Q.E.D.

In time series econometrics there are quite a few cases where we need a UWLLN for functions $g(.,\theta)$ depending on Z_{t-j} for all $j \geq 0$. In that case $g(.,\theta)$ takes a more general form as a random function:

Definition 3: *Let* $\{\Omega,\mathcal{F},P\}$ *be the probability space. A random function* $f(\theta)$ *on a subset* Θ *of a Euclidean space is a mapping* $f(\omega,\theta)$: $\Omega\times\Theta \to \mathbb{R}$ *such that for each Borel set B in* $\mathbb{R}$ *and each* $\theta \in \Theta$, $\{\omega \in \Omega: f(\omega,\theta) \in B\} \in \mathcal{F}$.

Definition 4: *A random function* $f(\theta)$ *on a subset* Θ *of a Euclidean space is almost surely continuous on* Θ *if there exists a set A with probability one such that for each* $\omega \in A$, $f(\omega,\theta)$ *is continuous in* $\theta \in \Theta$.

For example, let $Z_t \in \mathbb{R}$ be a stationary Gaussian moving average process or order 1 [alias an MA(1) process]:

$$Z_t = U_t - \alpha_0 U_{t-1},\ |\alpha_0| < 1,\ U_t \sim i.i.d.\ N(0,\sigma_0^2). \tag{10}$$

Then backwards substitution of $U_t = \alpha_0 U_{t-1} + Z_t$ yields $U_t = \sum_{j=0}^{\infty}\alpha_0^j Z_{t-j}$, hence

$$Z_t = -\sum_{j=1}^{\infty}\alpha_0^j Z_{t-1} + U_t \tag{11}$$

Thus, denoting $\mathscr{F}_t = \sigma(U_t, U_{t-1}, U_{t-2}, ...)$, the distribution of Z_t conditional on $\mathscr{F}_{t-1}$ is normal with conditional expectation $-\sum_{j=1}^{\infty}\alpha_0^j Z_{t-1}$ and conditional variance σ_0^2.

If the Z_t's were observable for all $t \leq n$, a version of the log-likelihood would take the form $\sum_{j=1}^{n} g_t(\theta)$, where

$$g_t(\theta) = -\frac{1}{2\sigma^2}\left(\sum_{j=0}^{\infty}\alpha^j Z_{t-j}\right)^2 - \frac{1}{2}\ln(\sigma^2) - \ln\left(\sqrt{2\pi}\right),\quad \theta = (\alpha,\sigma^2)^T, \tag{12}$$

is a random function. In that case we need to reformulate Theorem 3 as follows.

Theorem 4: *Let* $\mathscr{F}_t = \sigma(U_t, U_{t-1}, U_{t-2}, ...)$, *where* U_t *is a time series process with vanishing memory. Let* $g_t(\theta)$ *be a sequence of a.s. continuous random function on an open subset* Θ_0 *of a Euclidean space, and let* Θ *be a compact subset of* Θ_0. *If for each* $\theta_* \in \Theta$ *there exists an arbitrarily small* $\delta > 0$ *such that*

(*a*) $g_t(\theta_*)$, $\sup_{\|\theta-\theta_*\|\leq\delta} g_t(\theta)$ *and* $\inf_{\|\theta-\theta_*\|\leq\delta} g_t(\theta)$ *are measurable* $\mathscr{F}_t$ *and strictly stationary,*

(*b*) $E[\sup_{\|\theta-\theta_*\|\leq\delta} g_1(\theta)] < \infty$, $E[\inf_{\|\theta-\theta_*\|\leq\delta} g_1(\theta)] > -\infty$,

then $\operatorname{plim}_{n\to\infty}\sup_{\theta\in\Theta}|(1/n)\sum_{t=1}^{n} g_t(\theta) - E[g_1(\theta)]| = 0$.

2. *Consistency of M-estimators*

Theorems 3 and 4 are important tools for proving consistency of parameter estimators. A large class of estimators are obtained by maximizing or minimizing an objective function of the form $(1/n)\sum_{t=1}^{n} g_t(\theta)$, for example maximum likelihood estimators or nonlinear least squares estimators. These estimators are called M-estimators (where the M indicates that the estimator is obtained by Maximizing or Minimizing a Mean of random functions).

Suppose that the conditions of Theorem 4 are satisfied, and that the parameter vector of interest is

$$\theta_0 = \text{argmax}_{\theta \in \Theta} E[g_1(\theta)]. \tag{13}$$

Note that "argmax" is a short-hand notation for the argument for which the function involved is maximal. Then it seems a natural choice to use

$$\hat{\theta} = \text{argmax}_{\theta \in \Theta} (1/n)\sum_{t=1}^{n} g_t(\theta) \tag{14}$$

as an estimator of θ_0. Indeed, under some mild conditions the estimator involved is consistent:

Theorem 5: (*Consistency of M-estimators*) *Let* $\hat{\theta} = \text{argmax}_{\theta \in \Theta} \hat{Q}(\theta)$ *and* $\theta_0 = \text{argmax}_{\theta \in \Theta} \overline{Q}(\theta)$, *where* $\hat{Q}(\theta) = (1/n)\sum_{t=1}^{n} g_t(\theta)$ *and* $\overline{Q}(\theta) = E[\hat{Q}(\theta)] = E[g_1(\theta)]$. *If* θ_0 *is unique then under the conditions of Theorem 4,* $\text{plim}_{n\to\infty} \hat{\theta} = \theta_0$.

Proof: Since a continuous function on a compact set takes its maximum value in this set [see, for example, Bierens (2004, Appendix II)], it follows that $\hat{\theta} \in \Theta$ and $\theta_0 \in \Theta$. Moreover, by the same result it follows from the continuity of $\overline{Q}(\theta)$ and the uniqueness of θ_0 that for every $\varepsilon > 0$ for which the set $\{\theta \in \Theta: \|\theta - \theta_0\| \geq \varepsilon\}$ is non-empty,

$$\overline{Q}(\theta_0) > \sup_{\theta \in \Theta,\ \|\theta - \theta_0\| \geq \varepsilon} \overline{Q}(\theta) \tag{15}$$

[*Exercise*: Why?] Now by the definition of $\theta_{0,}$

$$0 \leq \bar{Q}(\theta_0) - \bar{Q}(\hat{\theta}) = \bar{Q}(\theta_0) - \hat{Q}(\theta_0) + \hat{Q}(\theta_0) - \bar{Q}(\hat{\theta})$$
$$\leq \bar{Q}(\theta_0) - \hat{Q}(\theta_0) + \hat{Q}(\hat{\theta}) - \bar{Q}(\hat{\theta}) \leq 2.\sup_{\theta\in\Theta}|\hat{Q}(\theta) - \bar{Q}(\theta)|, \tag{16}$$

and it follows from Theorem 4 that the right-hand side of (16) converges in probability to zero. Thus:

$$\text{plim}_{n\to\infty} \bar{Q}(\hat{\theta}) = \bar{Q}(\theta_0). \tag{17}$$

Moreover, (15) implies that for arbitrary $\varepsilon > 0$ there exists a $\delta > 0$ such that $\bar{Q}(\theta_0) - \bar{Q}(\hat{\theta}) \geq \delta$ if $\|\hat{\theta} - \theta_0\| \geq \varepsilon$, hence

$$P\left(\|\hat{\theta} - \theta_0\| > \varepsilon\right) \leq P\left(\bar{Q}(\theta_0) - \bar{Q}(\hat{\theta}) \geq \delta\right). \tag{18}$$

Combining (17) and (18), the theorem under review follows. Q.E.D.

It is easy to verify that Theorem 5 carries over to the "argmin" case.

References

Bierens, H. J. (2004): *Introduction to the Mathematical and Statistical Foundations of Econometrics*, Cambridge University Press, Cambridge, U.K.

Jennrich, R. I. (1969): "Asymptotic Properties of Non-Linear Least Squares Estimators", *Annals of Mathematical Statistics* 40, 633-643.