Notes and Supplementary Readings

Time Series Econometrics 2ST111

UPPSALA
UNIVERSITET

Yukai Yang, PhD

Uppsala Universitet
Statistiska Institutionen
P.O. 513, 751 20 Uppsala
Sverige

yukai.yang@statistik.uu.se

Contents

1 Stationary ARMA processes

A stationary autogressive moving average process of order p, q, ARMA(p, q) is defined to be

$$\phi(L)(y_t - \mu) = \theta(L)\varepsilon_t, \tag{1.1}$$

where the lag polynomial $\phi(L) = 1 - \sum_{i=1}^{p} \phi_i L^i$ is stable and the lag polynomial $\theta(L) = \sum_{i=0}^{q} \theta_i L^i$ is absolutely summable. The error sequence $\{\varepsilon_t\}$ is a white noise process (zero mean, constant variance and no serial correlation, but not necessary independent through time). In most cases, it is pressumed to be Gaussian white noise, which implies the likelihood form, the ergodicity for every moment, and the independency through time. The random variable y_t in (1.1) has the invariant unconditional expectation μ.

A stable lag polynomial $\phi(L)$ implies that it can be factorized as follows:

$$\phi(L) = 1 - \sum_{i=1}^{p} \phi_i L^i = \prod_{i=1}^{p}(1 - \lambda_i L), \tag{1.2}$$

with $|\lambda_i| < 1$ for $i = 1, ..., p$. Consider the special case when $p = 1$, $\phi(L) = 1 - \phi_1 L = 1 - \lambda_1 L$ with $\phi_1 = \lambda_1$, and the stability implies that $|\phi_1| < 1$.

An absolutely summable lag polynomial $\psi(L) = \sum_{i=0}^{\infty} \psi_i L^i$ satisfies $\sum_{i=0}^{\infty} |\psi_i| < \infty$. Apparently the lag polynomial $\theta(L)$ is absolutely summable as it is a finite series. Actually it can be regarded as $\psi_i = \theta_i$ when $i \leq q$ and $\psi_i = 0$ when $i > q$, a special case of the infinite series.

We have shown in the slides that a stable lag polynomial can be inverted to an absolutely summable one in the following way:

$$\phi(L)^{-1} = \left(1 - \sum_{i=1}^{p} \phi_i L^i\right)^{-1} = \prod_{i=1}^{p}(1 - \lambda_i L)^{-1}; \tag{1.3}$$

provided that $|\lambda_i| < 1$, we have

$$(1 - \lambda_i L)^{-1} = 1 + \sum_{j=1}^{\infty} \lambda_i^j L^j, \tag{1.4}$$

which satisfies $1 + \sum_{j=1}^{\infty} |\lambda_i^j| = (1 - |\lambda_i|)^{-1} < \infty$ (absolutely summable); and thus,

$$\phi(L)^{-1} = \prod_{i=1}^{p} \left(1 + \sum_{j=1}^{\infty} \lambda_i^j L^j\right) = \sum_{i=0}^{\infty} \tilde{\psi}_i L^i, \tag{1.5}$$

with $\tilde{\psi}_0 = 1$ and $\sum_{i=0}^{\infty} |\tilde{\psi}_i| < \infty$ due to that the product of absolutely summable lag polynomials are still absolutely summable.

A lag polynomial is invertible if its inverse exists. The stable lag polynomial is invertible, as its inverse is given by (1.5).

However, note that an absolutely summable lag polynomial may not be invertible. Consider the case when $q = 1$, $1 + \tilde{\theta}_1 L$ and $\tilde{\theta}_1 > 1$, see the example in Hamilton (1994) for more details.

The absolutely summable lag polynomial that we obtain by inverting the stable one is invertible, because, from (1.5), clearly we have

$$\left(\sum_{i=0}^{\infty}\tilde{\psi}_i L^i\right)^{-1} = \phi(L). \tag{1.6}$$

We presume that the absolutely summable lag polynomial $\theta(L)$ is invertible.

Hence, the following lag polynomial

$$\psi(L) = \sum_{i=1}^{\infty}\psi_i L^i = \phi(L)^{-1}\theta(L) \tag{1.7}$$

is invertible and absolutely summable.

Now let us consider the ARMA process (1.1) again. It follows that

$$\phi(L)(y_t - \mu) = \phi(L)y_t - \phi(L)\mu = \phi(L)y_t - c = \theta(L)\varepsilon_t,$$

where the intercept $c = \phi(1)\mu$, and $\phi(1) = 1 - \sum_{i=1}^{p}\phi_i$ (this is how the intercept comes). By rearranging and multiplying $\phi(L)^{-1}$ on both sides, we obtain

$$y_t = \phi(L)^{-1}c + \phi(L)^{-1}\theta(L)\varepsilon_t = \mu + \psi(L)\varepsilon_t, \tag{1.8}$$

which is exactly the Wold's decomposition. y_t has time-invariant expectation μ. The absolute summability of $\psi(L)$ ensures that y_t has finite unconditional variance, which is also time-invariant. The autocovariance structure γ_j can be obtained, see the slides or Hamilton (1994). It turns out that the autocovariances do not depend on time. Therefore, the ARMA process (1.1) is covariance-stationary.

From ARMA(p, q) to MA(∞) if we multiply the both sides of (1.1) by $\phi(L)^{-1}$, which is an absolutely summable lag polynomial, this yields a stationary moving average process of infinite order:

$$y_t - \mu = \phi(L)^{-1}\theta(L)\varepsilon_t = \psi(L)\varepsilon_t, \tag{1.9}$$

where $\psi(L)$ is an absolutely summable lag polynomial with infinite terms.

From ARMA(p, q) to AR(∞) if we multiply the both sides of (1.1) by $\theta(L)^{-1}$, this yields a stationary autoregressive process of infinite order:

$$\theta(L)^{-1}\phi(L)(y_t - \mu) = \varepsilon_t, \tag{1.10}$$

where $\theta(L)^{-1}\phi(L)$ is actually the inverse of $\psi(L)$.

Same rules hold for the conversions from AR(p) to MA(∞) and from MA(q) to AR(∞).

2 Maximum likelihood and the three tests

Consider the statistical model for the data X of sample size T from a sample space S. For example, in univariate case (the dependent variable is a scalar), the set X stacks the sequence of observations $\{y_t\}_{t=1}^{T}$, i.e., $X = (y_1, y_2, ..., y_T)$. Each observation y_t is a realization of its random variable and takes a value from its support $y_t \in S_t$. Then the sample space $S = S_1 \otimes S_2 \otimes ... \otimes S_T$.

The joint density function of X given the set of the parameters θ (any values of θ) is denoted by $p(X|\theta)$. The parameters take the values in the space Θ, i.e., $\theta \in \Theta$. Denote m the dimension of the parameter set (number of parameters).

Suppose that $\theta_0 \in \Theta$ is the set of the true parameters, from which the observations are drawn. The true data generating process has the joint density function $p(X|\theta_0)$. Unfortunately I do not know the value of θ_0.

2.1 maximum likelihood

Given the observations X, the likelihood function is $L(\theta) = p(X|\theta)$ for $\theta \in \Theta$, which is a function of θ. The maximum likelihood (ML) estimator is defined as

$$\hat{\theta} = \arg\max_{\theta \in \Theta} L(\theta), \tag{2.1}$$

or equivalently,

$$\hat{\theta} = \arg\max_{\theta \in \Theta} \log L(\theta), \tag{2.2}$$

where $\log L(\theta)$ is the logarithm of the likelihood, or simply log-likelihood function. Normally the log-likelihood function (2.2) is easier to handle as it takes the summation form, if the joint density belongs to the exponential family.

Consider the following AR(1) model

$$y_t = \phi y_{t-1} + \varepsilon_t, \tag{2.3}$$

for $t = 1, ..., T$ with y_0 given, where $\varepsilon_t \overset{iid}{\sim} N(0, \sigma^2)$. The data set is $X = (y_1, y_2, ..., y_T)$, and the parameter set is $\theta = (\phi, \sigma^2)$. The corresponding likelihood function is

$$L(\theta) = p(X|\theta) = \prod_{t=1}^{T} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{(y_t - \phi y_{t-1})^2}{2\sigma^2} \right\}. \tag{2.4}$$

The log-likelihood function is

$$\log L(\theta) = -\frac{T}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{t=1}^{T}(y_t - \phi y_{t-1})^2. \tag{2.5}$$

We see that, if the first order condition is applied here to find the optimum, the log-likelihood (2.5) is much easier to handle than the likelihood (2.4).

It is worth noting that in practice people solve the following the maximization problem in order to find the ML estimate.

$$\hat{\theta} = \arg\max_{\theta \in \Theta} \log L(\theta)/T, \tag{2.6}$$

where $\log L(\theta)/T$ is usually termed the average log-likelihood function.

One advantage of the average log-likelihood is to control the magnitude of the log-likelihood function. Note that the magnitude of the average log-likelihood is rather stable no matter how large the sample size T is. Moreover, the consistency of the ML estimator relies on whether the average log-likelihood will converge to some function with a unique global maximum with the maximizer being the true parameter.

2.2 hypothesis testing, score and information

A hypothesis concerning the parameter θ is expressed as a restriction on the parameter space Θ, i.e., the null hypothesis $H_0 : \theta \in \Theta_0 \subset \Theta$.

Consider the AR(1) model (2.3) whose parameter space is $\Theta = \Phi \otimes \Sigma$ where $\Phi = \mathbb{R}$ is the parameter space for ϕ and $\Sigma = \mathbb{R}^+$ for σ^2. A possible null hypothesis for the AR(1) model (2.3), for example, can be $\phi = 0$ or equivalently $\Theta_0 = \{0\} \otimes \Sigma$.

The ML estimator under the unrestricted model with $\theta \in \Theta$ is $\hat{\theta} = (\hat{\phi}, \hat{\sigma}^2)$. We denote $\tilde{\theta}$ the ML estimator under the restricted model with $\theta \in \Theta_0$.

$$\tilde{\theta} = \arg\max_{\theta \in \Theta_0} \log L(\theta), \tag{2.7}$$

For the above case, we have $\tilde{\theta} = (0, \tilde{\sigma}^2)$.

The score vector is defined to be the first-order derivative of the log-likelihood function

$$S_T(\theta) = \frac{\partial \log L(\theta)}{\partial \theta}. \tag{2.8}$$

It is a m-vector of functions of the parameter vector θ. People use very often the average score, which is defined as $\bar{S}_T(\theta) = S_T(\theta)/T$.

The information matrix is defined as follows:

$$I_T(\theta) = -\frac{\partial S_T(\theta)}{\partial \theta} = -\frac{\partial^2 \log L(\theta)}{\partial \theta^2}, \tag{2.9}$$

which is a $m \times m$ matrix. People use very often the average information, which is defined as $\bar{I}_T(\theta) = I_T(\theta)/T$.

Under the regularity conditions[1], the ML estimator $\hat{\theta}$ makes $S_T(\hat{\theta}) = 0$ which is exactly the first-order condition for maximization. And $I_T(\hat{\theta})$ must be positive definite as $\hat{\theta}$ is a maximizer.

The Fisher's information is defined as follows

$$\mathcal{I}_T(\theta) = \mathrm{E}_\theta[I_T(\theta)]. \tag{2.10}$$

E_θ reads the expectation with respect to θ. This mean that, given any $\theta \in \Theta$, for the fixed sample size T, we resample the data from the data generating process $p(X|\theta)$ for infinite times (population). For each sample, we compute its $I_T(\theta)$ and then take the average $\mathrm{E}_\theta[I_T(\theta)]$. Note

[1] If you are interested in the regularity conditions, please refer to the corresponding articles and books. We will not discuss about them here.

that the given θ for the data generating is known for this procedure, but we can change it for another procedure. Thus, in a formal representation, it is

$$\mathcal{I}_T(\theta) = \int_S I_T(\theta) p(X|\theta) \mathrm{d}X. \tag{2.11}$$

Still the Fisher's information is a function of θ. We input θ first, sample using the θ, and compute the function value $\mathcal{I}_T(\theta)$ at θ.

The information matrix $I_T(\theta)$ and the Fisher's information $\mathcal{I}_T(\theta)$ will both go to infinity when the sample size $T \to \infty$. However, the average information matrix and the average Fisher's information may converge to each other under certain conditions[2], i.e., $\bar{I}_\theta(\theta) \xrightarrow{p} \mathcal{I}(\theta)$ where $\mathcal{I}(\theta) = \lim_{T\to\infty} \mathcal{I}_T(\theta)/T$. This is the reason why sometimes one can replace $\mathcal{I}_T(\theta)$ by $I_T(\theta)$.

Theorem 1. *The score vector has the following properties:*

$$\begin{aligned} \mathrm{E}_\theta[S_T(\theta)] &= \mathbf{0}, & (2.12) \\ \mathrm{Var}_\theta[S_T(\theta)] &= \mathcal{I}_T(\theta). & (2.13) \end{aligned}$$

Proof. Obviously we have:

$$\begin{aligned} \int_S p(X|\theta)\,\mathrm{d}X &= 1 & (2.14) \\ \int_S \frac{\partial p(X|\theta)}{\partial \theta}\,\mathrm{d}X &= 0 & (2.15) \\ \int_S \frac{\partial^2 p(X|\theta)}{\partial \theta^2}\,\mathrm{d}X &= 0 & (2.16) \end{aligned}$$

Note that

$$S_T[\theta] = \frac{\partial \log p(X|\theta)}{\partial \theta} = p(X|\theta)^{-1} \frac{\partial p(X|\theta)}{\partial \theta}.$$

Then

$$\mathrm{E}_\theta[S_T(\theta)] = \int_S p(X|\theta)^{-1} \frac{\partial p(X|\theta)}{\partial \theta} p(X|\theta)\,\mathrm{d}X = \int_S \frac{\partial p(X|\theta)}{\partial \theta}\,\mathrm{d}X = 0.$$

And

$$I_T[\theta] = -\frac{\partial^2 \log p(X|\theta)}{\partial \theta^2} = -p(X|\theta)^{-1} \frac{\partial^2 p(X|\theta)}{\partial \theta^2} + \frac{\partial \log p(X|\theta)}{\partial \theta} \left(\frac{\partial \log p(X|\theta)}{\partial \theta} \right)'.$$

Thus,

$$\mathrm{E}_\theta[I_T(\theta)] = \mathrm{E}_\theta \left[\frac{\partial \log p(x|\theta)}{\partial \theta} \left(\frac{\partial \log p(x|\theta)}{\partial \theta} \right)' \right] = E_\theta[S_T(\theta) S_T(\theta)'] = \mathrm{Var}[S_T(\theta)].$$

□

The regularity conditions include the differentiability of the density function and the existence of the above integrals in the proof.

The theorem says that, if the observed data X with size T is a sample of draws from the model with the unknown true parameter θ_0, then the score $S_T(\theta_0)$ should have expectation 0 (around zero) and the covariance $\mathcal{I}_T(\theta_0)$. We will see in the following section that this is the basic idea of the score test.

[2] Note that the certain conditions are not identical to the regularity conditions.

2.3 the three tests

In this section, our interest focuses on the hypothesis testing for the null hypothesis $H_0 : \theta \in \Theta_0$. We call the model without any restrictions on the parameters, i.e., $\theta \in \Theta$, *the unrestricted model*, and we call the model with restrictions on the parameters, i.e., $\theta \in \Theta_0$, *the restricted model*. As defined in the previous section, $\hat{\theta}$ is the ML estimator under the unrestricted model, and $\tilde{\theta}$ the ML estimator under the restricted model.

Let us say that the parameter set can be split into two parts $\theta = (\alpha, \beta)$, and the true values of the parameters are $\theta_0 = (\alpha_0, \beta_0)$. The ML estimators for them under the unrestricted model are then given by $\hat{\theta} = (\hat{\alpha}, \hat{\beta})$. Actually we are testing $H_0 : \alpha = \breve{\alpha}$ where the vector α has dimension k with $1 \leq k \leq m$, and hence the ML estimators under the restricted model are precisely $\tilde{\theta} = (\breve{\alpha}, \tilde{\beta})$.

Based on the log-likelihood function, there are three tests: Wald test, Lagrange-multiplier or score test, and likelihood ratio test.

The Wald test statistic takes the form as follows:

$$\hat{\Sigma} = \begin{pmatrix} \hat{\Sigma}_\alpha & \hat{\Sigma}_{\beta\alpha} \\ \hat{\Sigma}_{\beta\alpha} & \hat{\Sigma}_\beta \end{pmatrix}$$

$$Q_{\text{wald}} = (\hat{\alpha} - \breve{\alpha})' \hat{\Sigma}_\alpha (\hat{\alpha} - \breve{\alpha}), \tag{2.17}$$

where the $k \times k$ matrix $\hat{\Sigma}_\alpha$ is the α-corresponding main diagonal block in $I_T(\hat{\theta})$. Typically the Wald test compares Q_{wald} with the $\chi^2(k)$ distribution, yields the p-value and makes statistical inferences. This is the one-sided test version. One can make it more flexible in the sense that one just compares $\hat{\alpha}$ with the Gaussian distribution $N(\breve{\alpha}, \hat{\Sigma}_\alpha^{-1})$, or even computes t-ratios for each element of $\hat{\alpha}$, and then the test becomes a two-sided one.

After all, the Wald test tells the story that, if the null hypothesis is true, then

$$\hat{\alpha} \sim N(\breve{\alpha}, \hat{\Sigma}_\alpha^{-1}) \tag{2.18}$$

approximately, due to the fact that $\hat{\theta} \sim N(\theta_0, I_T(\theta_0)^{-1})$ approximately under certain conditions. Note that $I_T(\theta_0)^{-1}$ is the Cramér-Rao lower bound for $\text{Var}(\hat{\theta})$.

(2.18) is asymptotically valid for stationary ARMA models with white noise errors under regularity conditions. The central limit theorem ensures that $\sqrt{T}(\hat{\theta} - \theta_0)$ will converge to a zero-mean multivariate Gaussian distribution with covariance $\mathcal{I}(\theta_0)^{-1}$. Note that $\mathcal{I}(\theta_0)$ is the limit of the average Fisher's information. (2.18) also says that $\hat{\alpha} \xrightarrow{p} \alpha_0$, as $\mathcal{I}_T(\theta_0)$ will explode.

Suppose that the null hypothesis is true, i.e., $\breve{\alpha} = \alpha_0$. Since the ML estimator is consistent, $\hat{\theta} \approx \theta_0$, $\mathcal{I}_T(\hat{\theta}) \approx \mathcal{I}_T(\theta_0)$, and $I_T(\hat{\theta}) \approx \mathcal{I}_T(\hat{\theta})$ as explained in the previous section. Then we believe that (2.18) is true and (2.17) is approximately $\chi^2(k)$ distributed.

However, in many cases, (2.18) is incorrect and (2.17) is not $\chi^2(k)$ distributed. For example, when the true model is a random walk or contains a unit root, the ML estimator is still consistent, but $\sqrt{T}(\hat{\alpha} - \alpha_0)$ does not converge to a Gaussian distribution. Then (2.18) does not hold, and (2.17) is not χ^2 distributed any more. This mistake results in a poor size property for the test. For a certain case in your analysis or research, you need to check carefully. In the literature, a common way to fix the Wald test is to suggest a better distribution for (2.17) to improve the size.

The Lagrange-multiplier (LM) test statistic or the score test statistic takes the form as follows:

$$Q_{\mathrm{lm}} = S_T(\tilde{\theta})' I_T(\tilde{\theta})^{-1} S_T(\tilde{\theta}) = T \bar{S}_T(\tilde{\theta})' \bar{I}_T(\tilde{\theta})^{-1} \bar{S}_T(\tilde{\theta}). \tag{2.19}$$

Typically the LM test compares Q_{lm} with the $\chi^2(k)$ distribution, yields the p-value and makes the inferences. It is a one-sided test. It follows the fact that the score vector at the true parameters $S_T(\theta_0)$ should have zero mean and covariance $\mathcal{I}_T(\theta_0)$, which is absolutely correct under the regularity conditions. But the distribution of the score is definitely not Gaussian in finite sample cases $T < \infty$.

The Gaussianity, again, comes from the central limit theorem under certain conditions. So in practice, you have to check carefully. If (2.19) is not asymptotically Gaussian distributed, then one needs to suggest another distribution for better size property.

Denote $S_\alpha(\theta) = \partial \log L(\theta)/\partial\alpha$ and $S_\beta(\theta) = \partial \log L(\theta)/\partial\beta$. If $S_\beta(\tilde{\theta}) = 0$ which is the first-order condition for the ML estimation of the restricted model, the LM test can be simplified to

$$Q_{\mathrm{lm}} = S_\alpha(\tilde{\theta})' \tilde{\Sigma}_\alpha^{-1} S_\alpha(\tilde{\theta}) = T \bar{S}_\alpha(\tilde{\theta})' \bar{\Sigma}_\alpha^{-1} \bar{S}_\alpha(\tilde{\theta}), \tag{2.20}$$

where $\bar{S}_\alpha$ is the average score, and $\tilde{\Sigma}_\alpha$ and $\bar{\Sigma}_\alpha$ are the α-corresponding main diagonal block in $I_T(\tilde{\theta})$ and $\bar{I}_T(\tilde{\theta})$, respectively.

Note that (2.19) and (2.20) are identical if $S_\beta(\tilde{\theta}) = 0$, but the equality does not hold when some nonlinear parameter enter, e.g., σ^2. In order to get rid of $S_{\sigma^2}(\tilde{\theta})$ in Q_{lm}, then you need to check the Fisher's information. You will find very soon that $I_T(\tilde{\theta})$ may be block diagonal with α-β off diagonal elements being zero, or that (2.19) and (2.20) are asymptotically equivalent if $\mathcal{I}(\theta_0)$ is block diagonal with α-β off diagonal elements being zero. Don't forget that, in principle, $\mathcal{I}_T(\theta_0)$ is the one that we should take for the LM test, which must be block diagonal.

The likelihood ratio (LR) test aims to compare how far the likelihood functions under the restricted and the unrestricted models are away to each other. If the ratio between them is close enough to one, then the null hypothesis is acceptable. The LR test takes the form as follows:

$$Q_{\mathrm{lr}} = -2 \log \left(\frac{L(\tilde{\theta})}{L(\hat{\theta})} \right) = 2 \log L(\hat{\theta}) - 2 \log L(\tilde{\theta}). \tag{2.21}$$

As $\Theta_0 \subset \Theta$, $L(\tilde{\theta}) \leq L(\hat{\theta})$, where the equality holds when $\tilde{\theta} = \hat{\theta}$ given that $\hat{\theta}$ is the unique global maximum of the likelihood function $L(\theta)$. This implies that $L(\tilde{\theta})/L(\hat{\theta}) \leq 1$ and $Q_{lr} \geq 0$. If the two likelihood functions are close enough, we can accept the null hypothesis, which means that this is a one-sided test.

Under certain conditions (you need to check each time), the LR test statistic can be compared with the $\chi^2(k)$ distribution. Typically people compute the p-value of the LR test for the hypothesis testing based on the χ^2 distribution, but confess that there may be size-distortion.

In the slides, we have shown that the three tests are asymptotically equivalent. We discuss it in more details but not rigorously in the following.

$$\frac{\xi}{\|\tilde{\theta}-\hat{\theta}\|^2} \xrightarrow{p} 0$$

First let us apply the Taylor expansion on the log-likelihood function at the estimator under the restricted model $\log L(\tilde{\theta})$. We expand it around the estimator under the unrestricted model. If the null hypothesis is true, then the two estimators should be asymptotically identical, which means that the expansion is valid and the approximation is accurate.

$$\log L(\tilde{\theta}) = \log L(\hat{\theta}) + (\tilde{\theta}-\hat{\theta})'S_T(\hat{\theta}) - \frac{1}{2}(\tilde{\theta}-\hat{\theta})'I_T(\hat{\theta})(\tilde{\theta}-\hat{\theta}) + \underline{o(||\tilde{\theta}-\hat{\theta}||^2)}^{=\ \xi} \tag{2.22}$$

where $o(||\tilde{\theta}-\hat{\theta}||^2)$ is the term that converges to zero faster than $||\tilde{\theta}-\hat{\theta}||^2$, i.e., $o(||\tilde{\theta}-\hat{\theta}||^2)/||\tilde{\theta}-\hat{\theta}||^2 \to 0$ as $||\tilde{\theta}-\hat{\theta}|| \to 0$, where $||\cdot||$ is a vector norm.

Note that $S_T(\hat{\theta}) = 0$ (for σ^2, consider the concentrated log-likelihood). Then we have

$$-2(\log L(\tilde{\theta}) - \log L(\hat{\theta})) = (\hat{\theta}-\tilde{\theta})'I_T(\hat{\theta})(\hat{\theta}-\tilde{\theta}) + o(||\hat{\theta}-\tilde{\theta}||^2), \tag{2.23}$$

where the left-hand side is the LR test, and the right-hand side is the Wald test. Note that the right-hand side is not the exact form of the Wald test that we introduced before. The Wald test only use $\hat{\alpha}$ and $\breve{\alpha}$ in the formula. Consider that if you are testing $H_0 : \theta = \breve{\theta}$ with $k = m$, then the right-hand side becomes exactly the Wald test.

Let us again expand $S_T(\hat{\theta})$ around $\tilde{\theta}$, and we obtain

$$S_T(\hat{\theta}) = 0 = S_T(\tilde{\theta}) - I_T(\tilde{\theta})(\hat{\theta}-\tilde{\theta}) + o(||\hat{\theta}-\tilde{\theta}||), \tag{2.24}$$

and hence

$$S_T(\tilde{\theta}) = I_T(\tilde{\theta})(\hat{\theta}-\tilde{\theta}) + o(||\hat{\theta}-\tilde{\theta}||), \tag{2.25}$$

We have the relations:

$$S_T(\tilde{\theta})'I_T(\tilde{\theta})^{-1}S_T(\tilde{\theta}) = (\tilde{\theta}-\hat{\theta})I_T(\hat{\theta})(\tilde{\theta}-\hat{\theta}) + o(||\hat{\theta}-\tilde{\theta}||). \tag{2.26}$$

where the left-hand side is the LM test, and the right-hand side is the Wald test.

Remember that, if the null hypothesis is true, $\breve{\alpha} = \alpha_0$ and both $\hat{\theta}$ and $\tilde{\theta}$ converge to θ_0 in probability. The distribution of all the three tests, under certain conditions, converges to $\chi^2(k)$.

This is not a rigorous proof for the asymptotic equivalence of the three tests, but it gives you the clue about how they are related to each other.

Though the three tests are asymptotically equivalent, they are totally different when you conduct them. In order to produce the Wald test statistic, you estimate the unrestricted model and get the estimates $\hat{\theta}$ and the information $I_T(\hat{\theta})$. For the LM test, instead of the unrestricted model, you estimate the restricted model and get the estimates $\tilde{\theta}$, the score $S_T(\tilde{\theta})$ and the information $I_T(\tilde{\theta})$. For the LR test, you have to estimate both the restricted model and the unrestricted model, compute the log-likelihood functions for both of them.

In practice, you may face the problem: "which test shall I choose?". It depends on many things. The convenience is one of the main reasons. If the unrestricted model is much more difficult to estimate than the restricted one, then you may probably consider the LM test. Sometimes the LR test may offer great convenience for derivation and the following statistical inference, though you have to estimate both models.

3 Asymptotic distribution theory

We make a summary of the asymptotic distribution theory here. First of all, in this section, we are talking about the sequence of random numbers, i.e., $\{X_t\}_{t=1}^{\infty}$. There are infinite random numbers in the sequence, from $t = 1$ to infinity. Each random number X_t follows its own distribution, may not be identical.

3.1 convergence in probability

We say that the sequence of the random variables X_t converges in probability to a constant c, if for any (small) positive number ε and any (small) positive δ, there exists a (big) integer N, such that for any $t \geq N$,

$$\text{Prob}\{|X_t - c| > \delta\} < \varepsilon, \tag{3.1}$$

A.S. Convergence: $\mathbb{P}\{\lim_{t\to\infty} X_t = c\} = 1$

$\Updownarrow$

$\mathbb{P}\{\bigcap^{\infty} \bigcup^{\infty} |X_t - c| > \delta\} = 0$

or, equivalently,

$$\lim_{t\to\infty} \text{Prob}\{|X_t - c| > \delta\} = 0, \tag{3.2}$$

Since each X_t has its own distribution, $\text{Prob}\{|X_t - c| > \delta\}$ implies a sequence of probabilities given δ. So the convergence in probability simply means that this sequence of probabilities converges to zero.

The convergence in probability can be represented in two ways:

$$\operatorname*{plim}_{t\to\infty} X_t = c, \tag{3.3}$$

$X_t \sim o_p(1) \iff X_t \xrightarrow{p} 0$

$X_t \sim O_p(t^n)$

$Z_t = \frac{X_t}{t^n}$

in which you can simply skip $t \to \infty$, or equivalently,

$$X_t \xrightarrow{p} c. \tag{3.4}$$

Consider if there are two sequences of random variables X_t and Y_t. Then $X_t \xrightarrow{p} Y_t$ simply implies that $X_t - Y_t \xrightarrow{p} 0$. Note that this does not mean that they converge in probability to the same constant. Instead they converge to the same random variable.

We say that an estimator $\hat{\theta}$ is consistent if $\hat{\theta} \xrightarrow{p} \theta_0$ where θ_0 is the true value of the parameter. In this case, we have hidden the subscript T (the sample size of the data) in $\hat{\theta}$.

3.2 convergence in mean square

The sequence of the random variables X_t is said to converge in mean square to a constant c, if for any (small) positive number ε, there exists a (big) integer N, such that for any $t \geq N$,

$$\text{E}(X_t - c)^2 < \varepsilon, \tag{3.5}$$

or, equivalently,

$$\lim_{t\to\infty} \text{E}(X_t - c)^2 = 0. \tag{3.6}$$

$\mathrm{E}(X_t - c)^2$ implies a sequence of these mean squares (moments).

The convergence in mean square can be represented as follows:

$$X_t \xrightarrow{m.s.} c. \tag{3.7}$$

It can also be called convergence in quadratic mean. It can be shown (Proposition 7.2 in Hamilton) that

$$X_t \xrightarrow{m.s.} c \implies X_t \xrightarrow{p} c, \tag{3.8}$$

while the other way around does not hold. And you need to know that, for a sequence of *i.i.d.* random numbers, its sample mean converges in mean square to its expectation.

3.3 convergence in distribution

Consider a sequence of random variables X_t for $t = 1, ..., \infty$, with the cumulative distribution functions (cdf) $F_t(x)$ for $t = 1, ..., \infty$. Suppose that there exists a random variable X with the cdf $F(x)$ such that

$X_t \xrightarrow{p} X \Rightarrow X_t \xrightarrow{d} X$

$$\lim_{t\to\infty} F_t(x) = F(x) \tag{3.9}$$

$X_t(\omega) \xrightarrow{p} X(\omega)$ $X_t(\omega) \to X(\delta).$

at any value x where $F(\cdot)$ is continuous. Then X_t is said to converge in distribution to X,

$$X_t \xrightarrow{d} X. \tag{3.10}$$

If you regard a constant c as a random variable with zero variance, you can say $X_t \xrightarrow{p} c \iff X_t \xrightarrow{d} c$.

In Proposition 7.3 (a) in Hamilton, it is true that

$$Y_t \xrightarrow{d} Y \text{ and } X_t - Y_t \xrightarrow{p} 0 \implies X_t \xrightarrow{d} Y. \tag{3.11}$$

However, it is NOT true that

$$Y_t \xrightarrow{d} Y \text{ and } X_t \xrightarrow{d} Y \implies X_t - Y_t \xrightarrow{p} 0. \tag{3.12}$$

The reason is that the convergence in distribution $\xrightarrow{d}$ only ensures that the limit distributions of X_t and Y_t have the same cdf, but they can be independent.

Proposition 7.3 (b) is in fact part of the Slutsky theorem. The complete version is as follows:

If $X_t \xrightarrow{p} c$ and $Y_t \xrightarrow{d} Y$, then

$$(1) \quad X_t + Y_t \xrightarrow{d} c + Y; \tag{3.13}$$

$$(2) \quad X_t Y_t \xrightarrow{d} cY; \tag{3.14}$$

$$(3) \quad Y_t / X_t \xrightarrow{d} Y/c. \tag{3.15}$$

The third one tells the story that if $X_t \xrightarrow{p} c$ then $1/X_t \xrightarrow{p} 1/c$.

3.4 small o and big o

In this section, we introduce the so-called "stochastic orders". The stochastic orders involves two notations: $o_p(\cdot)$ the small o and $O_p(\cdot)$ the big O.

For a sequence of random variables X_t, we say that $X_t = o_p(t^n)$ for some $n \in \mathbb{R}$ if $X_t/t^n \xrightarrow{p} 0$ as $t \to \infty$. For example, if $X_t \xrightarrow{p} 0$, then we say $X_t = o_p(1)$. Note that, in the literature, sometime people write $o_p(n)$ instead of $o_p(t^n)$.

For a sequence of random variables X_t, we say that $X_t = O_p(t^n)$ for some $n \in \mathbb{R}$ if for any (small) positive number ε, there exits a positive number (can be big) c and a positive integer N such that for any $t > N$

$$\text{Prob}\{|X_t|/t^n > c\} < \varepsilon \tag{3.16}$$

as $t \to \infty$.

Note that this definition is written in a different way in contrast to the definition of convergence in distribution. If $X_t = O_p(1)$, then X_t sequence is often referred to as "tight". Similarly, if $X_t = O_p(t^n)$, then $t^{-n}X_t$ is tight. Intuitively, the tightness means that the corresponding limiting distribution is somewhat "regular" in the sense that you can always find finite lower and upper bounds such that most of the probability "mass" is inside the bounded area, or "boundedness in probability".

We have some properties for the small o and big O:

- $X_t \xrightarrow{p} c \implies X_t = O_p(1)$
- $X_t \xrightarrow{d} X \implies X_t = O_p(1)$
- $X_t = op(t^n) \implies X_t = O_p(t^n)$
- $X_t = O_p(t^n) \implies X_t = op(t^m)$ if $m > n$
- if $X_t = O_p(t^n)$ and $Y_t = O_p(t^m)$, then $X_t + Y_t = O_p(t^{\max(n,m)})$ and $X_tY_t = O_p(t^{n+m})$, and the same results hold for $o_p(\cdot)$
- if $X_t = O_p(t^n)$ and $Y_t = o_p(t^m)$, then $X_tY_t = op(t^{n+m})$

3.5 central limit theorem

We have shown the central limit theorem for *i.i.d.* sequence in the previous lecture.

Proposition 7.4 in Hamilton has another name: the "delta method". We give the following theorem based on the small o and big O theory for free.

Theorem 2 (Stochastic Taylor Expansion). *Let X_t be a sequence of random variable in $\mathbb{R}^k$ with $X_t = c + O_p(t^n)$, where $c \in \mathbb{R}^k$ and $n < 0$, such that $t^n \to 0$ as $t \to \infty$. Then if f is continuously differentiable at c, we have*

$$f(X_t) = f(c) + f'(c)(X_t - c) + o_p(t^n) \tag{3.17}$$

Now we see that the well known "delta method" or Proposition 7.4 is a special case of it. But this theorem is more general.

4 Martingale difference sequence and central limit theorem

Definition 1 (Martingale). *A sequence of stochastic variables Y_t is called a martingale with respect to the information $\mathcal{F}_{t-1}$ available at time t, if Y_t has finite expectation and is measurable with respect to $\mathcal{F}_t$, and it holds that*

$$\mathrm{E}(Y_t|\mathcal{F}_{t-1}) = Y_{t-1}. \tag{4.1}$$

Intuitively, a martingale sequence has the property that the expected value of tomorrow's random variable is just today's observation, given all the information available today. Note that there is no requirements for any other moments, and even there is no need to be *i.i.d.* for martingale sequence. This offers great convenience for empirical data modelling and great robustness for the statistical inference. So does the martingale difference sequence as follows.

Definition 2 (Martingale difference sequence). *A sequence of stochastic variables X_t is called a martingale deference sequence with respect to the information $\mathcal{F}_t$ available at time t, if X_t has finite expectation and is measurable with respect to $\mathcal{F}_t$, and it holds that*

$$\mathrm{E}(X_t|\mathcal{F}_{t-1}) = 0. \tag{4.2}$$

You may ask why it has the name martingale difference sequence (MDS). Define $X_t = Y_t - Y_{t-1}$ where Y_t is martingale. Then clearly $\mathrm{E}(X_t|\mathcal{F}_{t-1}) = 0$. The MDS is the first order difference of a martingale.

Each element in the sequence of a MDS has unconditional zero mean due to

$$\mathrm{E}(X_t) = \mathrm{E}[\mathrm{E}(X_t|\mathcal{F}_{t-1})] = 0 \tag{4.3}$$

You can skip the "L^1-Mixingales" which is more general on pp.190, but please read the law of large number and the central limit theorem for the MDS.

In the following, we give another version of the vector MDS CLT, which may be used in your futher research.

Theorem 3 (Brown (1971)). *Let X_t, with finite variance, be a d-dimensional martingale difference sequence with respect to the information $\mathcal{F}_{t-1}$ available at time t. Assume that, as $T \to \infty$,*

$$T^{-1} \sum_{t=1}^{T} \mathrm{E}(\mathrm{X}_t \mathrm{X}_t' | \mathcal{F}_{t-1}) \xrightarrow{p} \mathbf{\Sigma}, \tag{4.4}$$

where $\mathbf{\Sigma}$ is positive definite. Assume further that, as $T \to \infty$, either

$$T^{-1} \sum_{t=1}^{T} \mathrm{E}\left[||\mathrm{X}_t||^2 \mathbf{1}\{||\mathrm{X}_t|| > \delta\sqrt{T}\} \;|\mathcal{F}_t \right] \xrightarrow{p} 0, \quad \text{or} \tag{4.5}$$

$$T^{-1} \sum_{t=1}^{T} \mathrm{E}\left[||\mathrm{X}_t||^2 \mathbf{1}\{||\mathrm{X}_t|| > \delta\sqrt{T}\} \right] \xrightarrow{p} 0 \tag{4.6}$$

hold for all $\delta > 0$. Then it holds that

$$T^{-1/2}\sum_{t=1}^{T}\mathbf{X}_t \xrightarrow{d} N_d(\mathbf{0},\mathbf{\Sigma}). \tag{4.7}$$

This theorem can also be found in Hall and Heyde (1980). Note that $||\cdot||$ is a matrix norm, and that $\mathbf{1}\{\cdot\}$ is an indicator function such that $\mathbf{1}\{\text{True}\} = 1$ and $\mathbf{1}\{\text{False}\} = 0$.

For better understanding, consider the example

$$X_t = \varepsilon_t Z_{t-1}, \tag{4.8}$$

where X_t is a scalar, the sequence ε_t is *i.i.d.* $(0,\omega)$, the sequence of scalars Z_t is stationary with zero mean and is ergodic in variance, i.e. $T^{-1}\sum_{t=1}^{T} Z_t^2 \xrightarrow{p} \mathrm{E}(Z_t^2) = \sigma^2$, and ε_t is independent of $(Z_{t-1}, \varepsilon_{t-1}, Z_{t-2}, ...)$. With

$$\mathcal{F}_t = \sigma(\varepsilon_i, Z_i), \quad i = 1, ..., t,$$

$\mathbf{X}_t$ is a MDS with respect to $\mathcal{F}_t$. Moreover,

$$T^{-1}\sum_{t=1}^{T}\mathrm{E}[X_t^2|\mathcal{F}_{t-1}] = \omega T^{-1}\sum_{t=1}^{T} Z_{t-1}^2 \xrightarrow{p} \omega\sigma^2. \tag{4.9}$$

Next, by the stationarity of X_t

$$T^{-1}\sum_{t=1}^{T}\mathrm{E}[X_t^2\mathbf{1}\{|X_t| > \delta\sqrt{T}\}] = \mathrm{E}[X_t^2\mathbf{1}\{|X_t| > \delta\sqrt{T}\}] \to 0 \tag{4.10}$$

as $T \to \infty$. The latter convergence holds by dominated convergence as $X_t^2\mathbf{1}\{|X_t| > \delta\sqrt{T}\} \le X_t^2$ and $\mathrm{E}(X_t^2) < \infty$. This shows that (4.4) and (4.6) hold and hence Theorem 3 gives

$$T^{-1/2}\sum_{t=1}^{T} X_t \xrightarrow{d} N(0,\omega\sigma^2). \tag{4.11}$$

References

Brown, B. M.: 1971, Martingale central limit theorems, *The Annals of Mathematical Statistics* **42**, 59–66.

Hall, P. and Heyde, C. C.: 1980, *Martingale Limit Theory and its Applications*, Academic Press, New York.

Hamilton, J. D.: 1994, *Time Series Econometrics*, Princeton University Press.