Lezione 10 - 22/10/2019

venerdì 1 novembre 2019 19:54

LEZIONE 10 - 22/10/2019

What we are going to do next week? Eview. We will use in our project. Show me first where we can find. Eview10 for students. The main limitation is in the volume of data. 2k data is enough for us. The other limitation has to due to the fact that we can't save our work so i can export it. The last thing: why this one? As we are going to see in a week time the most user friendly software for econometrics and statistics.

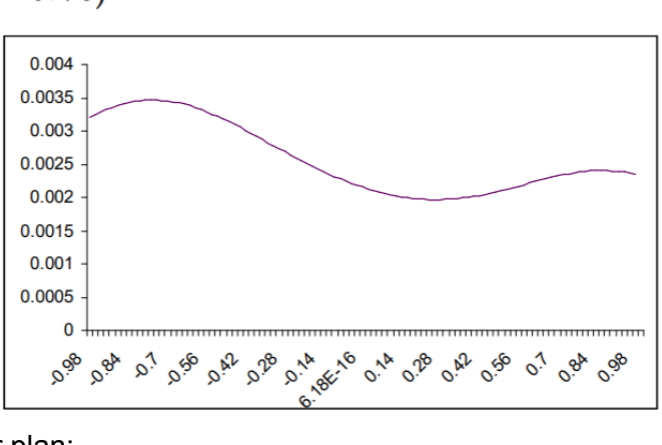
R is an alternative software that has the advantage to be more open to programming and require a big effort to interface with the software.

2 problems: invert matrix and …

The way to go is by using condition ad manage to break the product of all the marginal. So eliminating the huge matrix and get marginals.

Log-like hood resolve the problem of big matrix but generate the problem of generating all the value for FI. If we focus on the second line we can get a close formula and in the case of autoregression we solve problem of inverting a big matrix and computing the function for all the possible values and i got a solution to compute all the estimation of the values. We saw that for maximum value of MA i can use condition to manage to break the big inversion problem with a serious of factors but i still have to compute the density for all the possible value for theta. In the case of MA or model shaving MA (ARMA), the second part of problem (computing like-hood for all possible values) is still there. He shows that i can draw the function for all the possible values and i can actually draw it but is not a good way to proceeded.

The function may be computed for all the θ, $|\theta| < 1$ ($\hat{\theta} = -0.76$)



Another plan:

It became difficult to visualize. How do i think a 5 dimensional space? This is not the way to proceed. We will take any point, for example this one and i will go to the function and i got a point. I want to go to this maximum here. I have an idea: i can look at the function and see if it's going up or down (i look at the neighbours). I compute the derivative of the function that i want to maximize and i know i go up or down.

I know if i want to go to the right or the left for positive or negative increase. I want to go to the right. Next step will be to look at the cure itself. Here, in the last part he will give us a huge gain. When curve is flat i get the maximum or maybe the minimum. If curve is very stipe i have to take a great step. If is flat is not far away to we are looking for. Now i will go rightward and i want to take a step that will decide depending on the slop of the function. In this case i will take a small step and then repeat the exercise in the point. This procedure eventually, will get me from a point to another and eventually to the maximum. And cool thing is that i f get a small step because it will I adjust my step to the curvature. Looking at the first derivative and curvature (second derivative) i know where and how far i want to go. My starting point and ending point are pretty much the same. I can design an optimization process that get me there.

Optimisation of the objective function

In general, it is not always possible to obtain a closed form formula for the estimate, and it may be extremely time consuming to compute the log-likelihood function (even the conditional log-likelihood) for all the potential $\boldsymbol{\beta}$.

The optimisation of the log-likelihood may be carried using a numerical algorithm, such as the Newton-Raphson one.

Introduce

$$g(\boldsymbol{\beta}^{(0)}) = \left.\frac{\partial \mathcal{L}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(0)}} \quad \text{(gradient)}$$

$$H(\boldsymbol{\beta}^{(0)}) = -\left.\frac{\partial^2 \mathcal{L}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}'}\right|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(0)}} \quad \text{(Hessian)}$$

for a generic $\boldsymbol{\beta}^{(0)}$, and consider an approximate second order Taylor expansion of $\mathcal{L}(\boldsymbol{\beta})$,

$$\mathcal{L}(\boldsymbol{\beta}) \approx \mathcal{L}(\boldsymbol{\beta}^{(0)}) + \left[g(\boldsymbol{\beta}^{(0)})'\right][\boldsymbol{\beta} - \boldsymbol{\beta}^{(0)}] - \frac{1}{2}[\boldsymbol{\beta} - \boldsymbol{\beta}^{(0)}]' H(\boldsymbol{\beta}^{(0)})[\boldsymbol{\beta} - \boldsymbol{\beta}^{(0)}]$$

Recall that $\mathcal{L}(\boldsymbol{\beta})$ is maximised at $\hat{\boldsymbol{\beta}}$ if

$$\left.\frac{\partial \mathcal{L}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}} = 0.$$

Now, consider the approximation of the derivative around $\boldsymbol{\beta}^{(0)}$:

$$\frac{\partial \mathcal{L}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \approx [g(\boldsymbol{\beta}^{(0)})] - H(\boldsymbol{\beta}^{(0)})[\boldsymbol{\beta} - \boldsymbol{\beta}^{(0)}].$$

If the approximation was perfect, we could have just computed $\hat{\boldsymbol{\beta}}$ solving for $\boldsymbol{\beta}$

$$[g(\boldsymbol{\beta}^{(0)})] - H(\boldsymbol{\beta}^{(0)})[\boldsymbol{\beta} - \boldsymbol{\beta}^{(0)}] = 0,$$

i.e.,

$$\boldsymbol{\beta} = \boldsymbol{\beta}^{(0)} + H(\boldsymbol{\beta}^{(0)})^{-1}[g(\boldsymbol{\beta}^{(0)})].$$

However, this may be a rather poor estimate, because the approximation is not exact (there is a remainder, in this case of the third order, in the Taylor expansion of $\mathcal{L}(\boldsymbol{\beta})$). Let's call this possibly poor estimate $\boldsymbol{\beta}^{(1)}$, then, where

$$\boldsymbol{\beta}^{(1)} = \boldsymbol{\beta}^{(0)} + H(\boldsymbol{\beta}^{(0)})^{-1}[g(\boldsymbol{\beta}^{(0)})]:$$

clearly, this is (in a certain probabilistic sense) better than a generic $\boldsymbol{\beta}^{(0)}$.

If i get a point at the end, we will get the minimum. So, we have to be careful when we apply it. One thing is that this procedure works when we get a good starting value. Software will not for sure get us to the maximum.

Next, we can improve, by considering a second order approximation of $\mathcal{L}(\boldsymbol{\beta})$ in $\boldsymbol{\beta}^{(1)}$, and compute

$$\boldsymbol{\beta}^{(2)} = \boldsymbol{\beta}^{(1)} + H(\boldsymbol{\beta}^{(1)})^{-1}[g(\boldsymbol{\beta}^{(1)})].$$

The procedure can then be iterated until convergence (which gives $\hat{\boldsymbol{\beta}}$).

Example

ARMA(1,1) (assuming $\mu_0 = 0, \sigma_0^2 = 1$ known), $\boldsymbol{\beta} = (\theta, \phi)'$. Recall

$$\varepsilon_t(\boldsymbol{\beta}) = y_t - \phi y_{t-1} - \theta\varepsilon_{t-1}(\boldsymbol{\beta})$$

so for $t \geq 2$,

$$\frac{\partial \varepsilon_t(\boldsymbol{\beta})}{\partial \theta} = -\varepsilon_{t-1}(\boldsymbol{\beta}) - \theta\frac{\partial \varepsilon_{t-1}(\boldsymbol{\beta})}{\partial \theta}$$

$$\frac{\partial \varepsilon_t(\boldsymbol{\beta})}{\partial \phi} = -y_{t-1} - \theta\frac{\partial \varepsilon_{t-1}(\boldsymbol{\beta})}{\partial \phi}$$

$$\frac{\partial^2 \varepsilon_t(\boldsymbol{\beta})}{\partial \theta^2} = -2\frac{\partial \varepsilon_{t-1}(\boldsymbol{\beta})}{\partial \theta} - \theta\frac{\partial^2 \varepsilon_{t-1}(\boldsymbol{\beta})}{\partial \theta^2}$$

$$\frac{\partial^2 \varepsilon_t(\boldsymbol{\beta})}{\partial \phi^2} = -\theta\frac{\partial^2 \varepsilon_{t-1}(\boldsymbol{\beta})}{\partial \phi^2}$$

$$\frac{\partial^2 \varepsilon_t(\boldsymbol{\beta})}{\partial \theta\partial \phi} = -\frac{\partial \varepsilon_{t-1}(\boldsymbol{\beta})}{\partial \phi} - \theta\frac{\partial^2 \varepsilon_{t-1}(\boldsymbol{\beta})}{\partial \theta\partial \phi}$$

In many cases, you may start the optimisation with any set of starting values, but this may result in a rather slow optimisation, or even in an "incorrect" solution (you may end up picking a local maximum, rather than the maximum).
It is then advisable to start from a "good" point, that is, from a consistent estimate of $\boldsymbol{\beta}$ (tipically, an estimate that you may compute easily, even if it is less efficient than maximum likelihood): the correlogram based estimate is a good starting point (given certain regularity conditions, properties as in the pseudo-maximum likelihood estimate may be obtained after just one step).

Correlogram has a nice feature: i can compute the numbers so i can come out to a number at the end. The best way is taking estimate to the correlogram estimate and the solution will be in my neighbourhood.

Asymptotic Properties of parametric estimates

Correlogram base estimation and maximum like hood. At the end i will have and estimate. What we have to establish is weither this estimate is good or no. The general rule, there is no treason to say that estimation is good is estimate is maximum like hood. SO we have to study the properties of estimation.

Finally the most interesting case is the pseudo maximum likelihood: max estimation when we pretend to have a Normal distribution.

We want that estimate is consistent!

Limit properties: consistency

Then

$$\hat{\boldsymbol{\beta}} \to_p \boldsymbol{\beta}_0 \text{ as } T \to \infty$$

i.e. as $T \to \infty$, $\hat{\boldsymbol{\beta}}$ (any of $\hat{\boldsymbol{\beta}}_C$, or of $\hat{\boldsymbol{\beta}}_{ML}$ or of $\hat{\boldsymbol{\beta}}_{PML}$) is a consistent estimate of $\boldsymbol{\beta}_0$.

It also holds that $\hat{\sigma}_C^2 \to_p \sigma_0^2$, $\hat{\sigma}_{ML}^2 \to_p \sigma_0^2$ and $\hat{\sigma}_{PML}^2 \to \sigma_0^2$ as $T \to \infty$ (where $\hat{\sigma}_C^2$, $\hat{\sigma}_{ML}^2$ and $\hat{\sigma}_{PML}^2$ are the correlogram based, ML and PML estimates of σ_0^2, respectively).

The meaning requirement is having consistency and we can prove it.

Another properties is if we know distribution of the estimates. He can give us what is the estimate and how precise the estimate is. This estimate are all asymptotical normal and it's good because is very familiar for me.

Limit properties: asymptotic normality

$$\sqrt{T}\left(\hat{\boldsymbol{\beta}}_C - \boldsymbol{\beta}_0\right) \to_d N(0, \Sigma_C)$$

$$\sqrt{T}\left(\hat{\boldsymbol{\beta}}_{ML} - \boldsymbol{\beta}_0\right) \to_d N(0, \Sigma_{ML})$$

$$\sqrt{T}\left(\hat{\boldsymbol{\beta}}_{PML} - \boldsymbol{\beta}_0\right) \to_d N(0, \Sigma_{ML})$$

as $T \to \infty$, $\sqrt{T}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right)$ is asymptotically normally distributed. Notice however the dispersion is, in general, different.

The second part good is that i can derive variance covariance matrix. He will not give us the proof but he show us that i can put number in the matrix.

Examples of Σ_{ML}:

$$AR(1): \sqrt{T}\left(\hat{\phi} - \phi_0\right) \to_d N(0, 1 - \phi_0^2)$$

$$AR(2): \sqrt{T}\left(\begin{bmatrix} \hat{\phi}_1 \\ \hat{\phi}_2 \end{bmatrix} - \begin{bmatrix} \phi_{0;1} \\ \phi_{0;2} \end{bmatrix}\right) \to_d N\left(0, \begin{bmatrix} 1-\phi_{0;2}^2 & -\phi_{0;1}(1+\phi_{0;2}) \\ -\phi_{0;1}(1+\phi_{0;2}) & 1-\phi_{0;2}^2 \end{bmatrix}\right)$$

$$MA(1): \sqrt{T}\left(\hat{\theta} - \theta_0\right) \to_d N(0, 1 - \theta_0^2)$$

$$MA(2): \sqrt{T}\left(\begin{bmatrix} \hat{\theta}_1 \\ \hat{\theta}_2 \end{bmatrix} - \begin{bmatrix} \theta_{0;1} \\ \theta_{0;2} \end{bmatrix}\right) \to_d N\left(0, \begin{bmatrix} 1-\theta_{0;2}^2 & -\theta_{0;1}(1-\theta_{0;2}) \\ -\theta_{0;1}(1-\theta_{0;2}) & 1-\theta_{0;2}^2 \end{bmatrix}\right)$$

MA(1) $Y_t = \varepsilon_t + \theta\varepsilon_{t-1}$

MA(2) $Y_t = \varepsilon_t + \theta_1\varepsilon_{t-1} + \theta_2\varepsilon_{t-2}$ ↑ 0

ARMA(1,1) $Y_t = \phi Y_{t-1} + \varepsilon_t + \theta\varepsilon_{t-1}$ ↑ 0

$$ARMA(1,1): \sqrt{T}\left(\begin{bmatrix} \hat{\phi} \\ \hat{\theta} \end{bmatrix} - \begin{bmatrix} \phi_0 \\ \theta_0 \end{bmatrix}\right) \to_d N\left(0, \begin{bmatrix} (1-\phi_0^2)^{-1} & (1+\phi_0\theta_0)^{-1} \\ (1+\phi_0\theta_0)^{-1} & (1-\theta_0^2)^{-1} \end{bmatrix}^{-1}\right)$$

the last variance can be rewritten as

$$\frac{1+\phi_0\theta_0}{(\phi_0+\theta_0)^2} \times \begin{bmatrix} (1-\phi_0^2)(1+\phi_0\theta_0) & -(1-\theta_0^2)(1-\phi_0^2) \\ -(1-\phi_0^2)(1-\theta_0^2) & (1-\theta_0^2)(1+\phi_0\theta_0) \end{bmatrix}$$

★These do not depend on σ_0^2;
★The estimates in the AR(1), MA(1) are more precise the stronger the dependence.

The roof start from LLN slide but is not examinable but is useful to have in the lectures.

Not only i can put number, but i can compute. I can tell you exactly what is the estimate variance of this guy.

Which means that if we are interested in testing FI 0.5 we can do it. I can test hypothesis on this parameters.

Summarize

This estimates are consistent so it will match. The second proprieties is that is asymptotically normal and likely they are. Look at the formula of the variance.

We can see that the function are number that i can compute. If i can estimate this i can compute this functions.

This formula is of ML. He didn't' present variance for this functions. Bad news is that formula from correlogram is different form the ML estimates. In general, the CL estimates will not be good as the ML estimates. The second will be more precise and have less variance. One situation of CL good as ML is when i have autoregression. It's regression in both functions. ML > CL (more précises).

CL in the other hand has a value because he gets me to the right starting value. And that's why we are studying the CL.

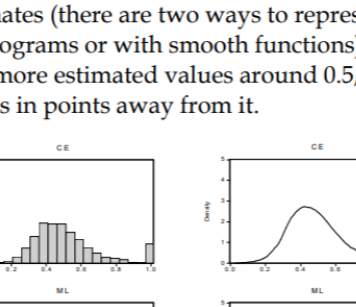
What does it mean that CL is not good as ML?

Properties of the Correlogram Based estimate and Maximum Likelihood estimate

What does it mean to say that the Maximum Likelihood estimate is more precise than the Correlogram based estimate?

✠ Example 1. MA(1).

The series



was generated as MA(1) with $\theta = 0.5$.

★ If we pretend not to know θ, and we estimate it as correlogram based or maximum likelihood estimate,

$$\hat{\theta}_C = 0.35, \hat{\theta}_{ML} = 0.43$$

so in this particular example $\hat{\theta}_{ML}$ got closer to θ (so, it worked better).

This is just one example and these are random variable. Instead of doing one example i do 1k example.

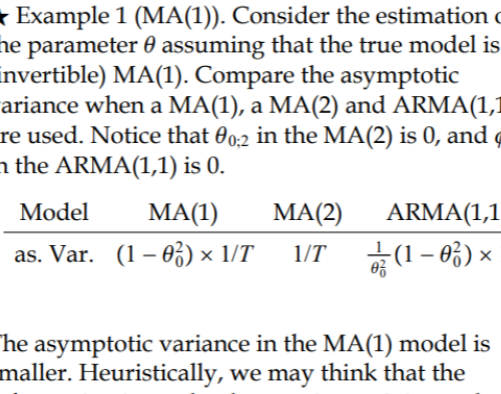
✠ Example 2. 1000s MA(1), an experiment.

I took 1000 random series from the same process:

★ the estimate $\hat{\theta}_{ML}$ gets closer to 0.5 than $\hat{\theta}_C$ does in 68.5% of the cases;

★ the standard error of the estimated values $\hat{\theta}_{ML}$ is 0.075, the standard error of the estimated values $\hat{\theta}_C$ is 0.104.

★ We can look at the whole sample distribution of the estimates (there are two ways to represent it, with histograms or with smooth functions). $\hat{\theta}_{ML}$ clusters more estimated values around 0.5, and much less in points away from it.



All this means that $\hat{\theta}_{ML}$ is more precise than $\hat{\theta}_C$ in a statistical sense.

Another way is looking at the standard error of the variable. The standard error of the estimates is 0.075 for ML and 0.104 for CL. So, a difference of 30%. If you want to understand what this 30 % mean we can see the distributions.

Interpretation of the standard errors and application to testing

The standard errors can be seen as a measure of the precision of the estimate, and can be also used in testing.

★ Example 1 (MA(1)). Consider the estimation of the parameter θ assuming that the true model is an (invertible) MA(1). Compare the asymptotic variance when a MA(1), a MA(2) and ARMA(1,1) are used. Notice that $\theta_{0;2}$ in the MA(2) is 0, and ϕ_0 in the ARMA(1,1) is 0.

Model	MA(1)	MA(2)	ARMA(1,1)
as. Var.	$(1-\theta_0^2) \times 1/T$	$1/T$	$\frac{1}{\theta_0^2}(1-\theta_0^2) \times 1/T$

The asymptotic variance in the MA(1) model is smaller. Heuristically, we may think that the information is used only to estimate θ, instead of dispersing it to estimate also θ_2 or ϕ.

ML is better in most cases.

$$\sqrt{T}\left(\hat{\boldsymbol{\beta}}_{PML} - \boldsymbol{\beta}_0\right) \to_d N(0, \Sigma_{ML})$$

It's not a typo. It's a huge meaning: in fact, knowing info are N distributed it's worth anything. We understand that the ML is just a convenient way to define an estimate. What we understood is the ML is a Conditional maximum like hood. When we press maximum like hood on the computer it's doing square (?). We don't have to worry about the distribution of the estimate. We will get the same estimate regardless of my distribution.

When i estimate AR(1)

Examples of Σ_{ML}:

$$AR(1): \sqrt{T}\left(\hat{\phi} - \phi_0\right) \to_d N(0, 1 - \phi_0^2)$$

$$AR(2): \sqrt{T}\left(\begin{bmatrix} \hat{\phi}_1 \\ \hat{\phi}_2 \end{bmatrix} - \begin{bmatrix} \phi_{0;1} \\ \phi_{0;2} \end{bmatrix}\right) \to_d N\left(0, \begin{bmatrix} 1-\phi_{0;2}^2 & -\phi_{0;1}(1+\phi_{0;2}) \\ -\phi_{0;1}(1+\phi_{0;2}) & 1-\phi_{0;2}^2 \end{bmatrix}\right)$$

$$MA(1): \sqrt{T}\left(\hat{\theta} - \theta_0\right) \to_d N(0, 1 - \theta_0^2)$$

$$MA(2): \sqrt{T}\left(\begin{bmatrix} \hat{\theta}_1 \\ \hat{\theta}_2 \end{bmatrix} - \begin{bmatrix} \theta_{0;1} \\ \theta_{0;2} \end{bmatrix}\right) \to_d N\left(0, \begin{bmatrix} 1-\theta_{0;2}^2 & -\theta_{0;1}(1-\theta_{0;2}) \\ -\theta_{0;1}(1-\theta_{0;2}) & 1-\theta_{0;2}^2 \end{bmatrix}\right)$$

This telling me how precise is the estimates.

Interpretation of the standard errors and application to testing

The standard errors can be seen as a measure of the precision of the estimate, and can be also used in testing.

★ Example 1 (MA(1)). Consider the estimation of the parameter θ assuming that the true model is an (invertible) MA(1). Compare the asymptotic variance when a MA(1), a MA(2) and ARMA(1,1) are used. Notice that $\theta_{0;2}$ in the MA(2) is 0, and ϕ_0 in the ARMA(1,1) is 0.

Model	MA(1)	MA(2)	ARMA(1,1)
as. Var.	$(1-\theta_0^2) \times 1/T$	$1/T$	$\frac{1}{\theta_0^2}(1-\theta_0^2) \times 1/T$

The asymptotic variance in the MA(1) model is smaller. Heuristically, we may think that the information is used only to estimate θ, instead of dispersing it to estimate also θ_2 or ϕ.

In these three cases the estimation is good. But we can compare the three possible way to estimate the model. MA(1) has the smallest variance and will be the best model of estimation.

Last application of standard error is for testing. And if we go back to the MA(1) : radT .. If i want to standardize this quantity what i nneed to do is to divide the second term to the rad of
T. And this is the standardize form.

✠ Example 2 (MA(1)).

Suppose that a MA(1) model is estimated (via ML/CML), with 100 observations, and $\hat{\theta}$ takes value 0.8.

The standard error, $\sqrt{\frac{1-\theta_0^2}{T}}$ is not observable (because we do not know θ_0). The estimate takes value $\sqrt{\frac{1-0.8^2}{100}} = 0.06$.

If we want to test $H_0 : \{\theta_0 = \theta\}$ we use

$$\sqrt{T}\frac{\left(\hat{\theta} - \theta_0\right)}{\sqrt{1-\theta_0^2}} \to_d N(0,1)$$

so for example, to test

$$H_0 : \{\theta_0 = 0.7\} \text{ vs } H_A : \{\theta_0 \neq 0.7\}$$

the test statistic under the null hypothesis takes value 1.4003, so the null hypothesis is not rejected.

Hip will not be rejected.

Or we can do the same with MA(2)

✠ Example 3 (MA(2)).

Suppose that a MA(2) model is estimated (via ML/CML), with 100 observations, and $\hat{\theta}_1$ takes value 0.8, $\hat{\theta}_2$ takes value 0.05.

The standard error, $\sqrt{\frac{1-\theta_{0;2}^2}{T}}$ is not observable (because we do not know $\theta_{0;2}$). The estimate takes value $\sqrt{\frac{1-0.05^2}{100}} = 09.9875$.

If we want to test $H_0 : \{\theta_{0;1} = \theta\}$ we use

$$\sqrt{T}\frac{\left(\hat{\theta} - \theta_{0;1}\right)}{\sqrt{1-\theta_{0;2}^2}} \to_d N(0,1)$$

Notice that this require knowledge of $\theta_{0;2}^2$, and this not know not even under H_0: we can, however, replace it by a consistent estimate ($\hat{\theta}_2$).

So for example, to test

$$H_0 : \{\theta_{0;1} = 0.7\} \text{ vs } H_A : \{\theta_{0;1} \neq 0.7\}$$

the test statistic under the null hypothesis takes value 1.0013, so the null hypothesis is not rejected.

Variance is Theta 0.2 squared. And then i can compute the statistics and get a hyp value that is 1.003

Or the same story in ARMA(1,1)

✠ Example 4 (ARMA(1,1)). Suppose that an ARMA(1,1) model is estimated (via ML/CML), with 100 observations, and $\hat{\phi}$ takes value 0.8, $\hat{\theta}$ takes value 0.05.

If we want to test $H_0 : \{\phi_0 = \phi, \theta_0 = \theta\}$ we use the Wald test statistic

$$T\begin{pmatrix} \hat{\phi} - \phi_0 & \hat{\theta} - \theta_0 \end{pmatrix} \times \left(\begin{bmatrix} (1-\phi_0^2)^{-1} & (1+\phi_0\theta_0)^{-1} \\ (1+\phi_0\theta_0)^{-1} & (1-\theta_0^2)^{-1} \end{bmatrix}^{-1}\right)^{-1} \times \begin{pmatrix} \hat{\phi} - \phi_0 \\ \hat{\theta} - \theta_0 \end{pmatrix} \to_d \chi_2^2$$

(i.e., the Wald test statistic is asymptotically χ_k^2 distributed, with k equal to the number of parameters being tested).

So for example, to test

$$H_0 : \{\phi_0 = 0.7, \theta_0 = 0.2\}$$

vs

$$H_A : \{\phi_0 \neq 0.7, \&/\text{or } \theta_0 = 0.2\}$$

the test statistic takes value 1.6730, so the null hypothesis is not rejected with size 5% (c.v. 5.99).

We can test more hypothesis at the same time but we are not testing just a scalar but a vector. If observation is N distributed and we obtain a vector to taking to account all the variance covariance matrix. The way we do it is by squaring everything.

Estimates are always consistent, distributed normally and is not important to know what is the distribution of observable staff.

How precise is an estimates with variance.

Finally variance can be used in testing.