

Lezione 8 - 15/10/2019

venerdì 1 novembre 2019 19:50

Lezione 8 – 15/10/2019

Chapter 5 again.

ARMA model are very versatile

They only require a small number of parameters and we have to estimate a small number of parameters.

We saw this technique were basically we estimate parameter simply by matching the moment. So first order of autocorrelation has the value of FI and idem for the sample value of the autocorrelation. Estimation from the correlogram, was invented by Walker, so it is also referred as Walker estimation. Some people have strictly approach and the estimate autoregression parameter. That's one way to estimate parameters.

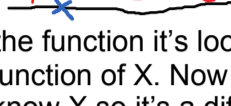
If you did some statistics or econometric you will know there is another way to estimate parameters as well. That's maximum **likelihood**

Jump to the appendix, as a review for someone who's does not have statistic background. Simplest case: imagine an experiment and may consider taking an exam and passing it. The probability is p . The probability for n student is $P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$

If we know p we can compute the probability of n persons. If we don't know p we can't compute that. This function is used when we know p and we want to find the probability that will observe some realization. This is the binomial function.

Let's make the reverse experiment: when I don't know p of each one, so I don't know p . 5 people pass the exams. What does it tell me about the true probability of passing? If I go to another class with 7 students maybe I will not have 5 people passing. So, it's informative but is not the real probability of passing. Then I ask myself, let's suppose the prob of passing is 0.4. Then I will observe 5 passes when probability is 0.077. If $p = 0.5$ I will see 0.275. I don't know p but I have some information about it. I still know p , but if I have to choose a value for p I will take 0.8 because it gave me the major probability. It's more likely to observe 0.275 than 0.077 to happen. Of course, there's not particular reason why I choose that values, I could take any number between 0 and 1. What does it looks like?

[Grafico]



If you look at the function it's looks like the function with I started with. But it's not, the binomial is a function of X . Now on the other hand this is the function of Θ and we think that we know X so it's a different function. This is a function of Θ not X . This way of changing the function transform the prob in a new thing that is called likelihood.

Where is the maximum θ ?

It's the one on 0.5. The purpose is that even if I have something that seems the same function is not the same function. This is a different thing; binomial only take value for 7. The likelihood will take value for value from 0 to 1 and this approach is called maximum likelihood.

I will rather thing that I see something that has 0.275 prob to happen that something with 0.077 prob to happened. So, it's called maximum likelihood to happen.

What he don't show is that I can draw a picture and find out that 5/7 is the maximum but we also like to establish this maximum mathematically. If I takes this function and obtain the maximum with derivative = 0.

This will be a familiar setting: I got a sample of X and Y where X is deterministic. So X could be time, cosine (coseno) function. The distribution of Y is not normal, and the variance is the variance of the normal component. So, we study it when we studied linear regression.

It is normally distributed, and we get the normal. It's $Y - \text{the mean}$. The product of all marginal is the join distribution.

The second one is a function in function of α , β and σ^2 . They look the same but they are not the same. Changes the observations.

It is actually a function of α and β . Y is not taken as known. At the end of the day we still don't know u and y . We swap the function and it's the likelihood. Join density is know transformed in join likelihood. Density is on the top of the slide, likelihood it on the bottom of the slide.

The maximum of likelihood. Instead of maximizing that function we take the logarithm of the likelihood that is a monotonic function. Then I will call my estimate as I called before: α , β , σ are estimations of these parameter. We can solve it mathematically.

The derivative of u respect to α will be 1 and the derivative of β is X . So we have twice the derivative of α and u .

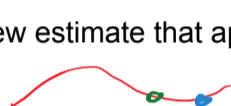
I have α and β unknown. I solve it and get estimation of α and β . Moreover, using the fact that deviation for the mean is adapt to 0 so β could be written even in this way (bottom of the slide).

If $\alpha = 0$ I can drop it and get simpler solution. Even I can compact in matrix.

Notice that $(x' x)^{-1}$ is a function of the observation. We will not get the value of β (as we didn't got the true probability of passing). As I change the observation I get a different observation for β . So β is itself a random variable with distribution that is a normal distribution.

This problem that we solve pointed out that the maximization of the likelihood to minimizing the function on the right of the logarithm. So we have to minimize the sums of $u(\alpha, \beta)^2$. This gives me a way to estimate α and β even without the information. It is so because we have the assumption that is not normally distributed. Beta function is a function of our observation and we could use the same function for something not normally distributed in this way. We generalize the estimation and we apply this function in general for estimation that minimize the sum of $u(a, b)^2$. This approach takes the function as the estimate, know that minimizing this value is called Minimum square even when function is not normally distributed.

I took the maximum estimation, and in this way I define a new estimate that apply even without the assumption of the normal distribution.



How does this concept works if I have a time series?

Before we get the likelihood and we obtain the maximum. Join density of the marginal because the observation are all independent distributed. If not? The join density will not look like this. We can actually write this. X , Y and the Covariance of X and Y . If I got n of them? It's difficult to write it down.

See this density: factors with $\text{rad}(2 \text{PI} \sigma^2)$. It's easy to divide by the variance because it's scalar. But when I have more observation what is the variance? The variance and autocovariance are part of the same story. If I think the observation as a vector, the variance will be the autocovariance of everything. The covariance will look like $(Y - u)'(Y - u)$. We assume that $u = 0$.

This function depend not only of variance but also the covariance of Y_n . The variance is actually a matrix so we called the autocovariance-variance vector. We stick this guys to some sort of join density.

The join density will be in general [slide Estimation maximum likelihood].

The square appear because the vector $y - u'$ is multiplied with $y - u$ vector. Omega variance???

I must think of this not as function of observation but function of parameter that I want to estimate.

MA(1) the autocovariance are $(1 + \theta^2)$, θ and 0 everywhere else. I will factor element outside. How θ fits in the matrix of Omega. I know what the value of θ is. If I'm interested in estimation the mean, μ is also an unknown. So, my parameter make feature here into the $\omega(b)$ matrix.

All parameter of interest are all parameter of ARMA model.

SOME COMMENTS USEFULL WHEN WE'LL REVISE [SLIDE]

[Strongly revise Hamilton that call θ in bold not β]

We will use 0 that is the true parameter that generate the data, the parameter we want to estimate

Example

I saw 4 point and I want to estimate θ . I have to do the Omega matrix. I take one potential value (0.5 for example) and I stick it to the likelihood function. Then we take other value for θ . **The estimation is -0.5 because is the maximum estimate value.**

In many case...

E' ESAMINABILE

BUT I don't like inverse of matrix. I have to do this inversion for all the point of θ NOT NICE. This is a 4×4 but really we could have 1000 of observation.

Another problem: this is a 1000 that I have to invert. If I estimated MA(2) I have to estimate 2 parameter so it's double infinity inversion. **THIS IS NOT GOING TO WORK.**

Maximum likelihood is an inspiration, not a something possible.

From now on that's how I will act in a fantasy world, let's see how I will go in a real world.

Example of AR(1).

For every single observation we begin that are normally distributed.

The density will be this guy [SLIDE]

Let's think of Y_2 . I could also use that the join density is the conditional density * times the marginal and $f_{Y_2|Y_1}$. Y_1 is threatened as we now it because is not random.

I can break the join density in the product of all conditional * the marginal. Y_{t-1} is the only interesting. To turn this density in likelihood we express for of Y .

The maximum value of inverting the matrix is turned in a simple minimization. Is this still maximum value? YES. As I look at it, there something special. The first line is the density of the first observation and second to the point 2 all the way to T . I see this likelihood in the AR example before. Instead of Y_t we called it u_t . So, I know how to solve it and if I do all the step I know that is my estimate. So, I don't actually have to compute likelihood in every possible point of T . It does not work for a bit. The maximum of T (poi boh non ho capito).

Let's not invent a new estimate and forget about the In line and work on the second. What does it mean to do that?

The problem of the second line as a density. Everything given Y_1 . So really If I forget about the first line. I do a maximum likelihood in a different design. Y_1 is not random but it's given. This guy he proposes us is a different likelihood with a different design. Y_1 is not random. We called it **conditional maximum likelihood**. (it's different form the maximum likelihood).

When I have bigger autoregression (Big lags)? I will have the same thing. Join density is the product of all the marginal. If I have MA(1) one period, MA(p) I have p periods, so I will general condition on p dimensional density.

Basically, start form maximum likelihood and take away a bit of Maximum Likelihood I got something similar but simpler to compute.

The estimate of AR is the estimation.

For MA(1) I will get the same arguments. If I have MA(1) what I did before was condition on the past (E_{p-1}). If I conditional on E_{p-1} given the density of the normal. The variance of this guys will be in function of eps. The big difference in the AR(1) I was conditioning on the Y_{t-1} . This is something that I observed. But I don't observe E_{p-1} . Maybe I can compute it? **How do I compute E_{p-1} ? I could compute looking at his previous values.**

For example

I would like to use: $Y_t = \text{eps}_t + \theta \text{eps}_{t-1}$. I want to get E_{p-1} .

Let's look at $Y_{t-1} = \text{eps}_{t-1} + \theta \text{eps}_{t-2}$

$E_{p-1} = Y_{t-1} - \theta \text{eps}_{t-2}$

The problem is that I trade E_{p-1} with E_{p-2} that I don't know.

So $Y_t = E_{p-1} + \theta Y_{t-1} - \theta^2 \text{eps}_{t-2}$.

I will have eps_{t-2} in way and it's not theta. How do I go by avoiding this problem?

If I have 1000 observation I can go backward as I want. I want to go back all the way to Y_1 and we get eps_0 that we don't have. But now let's pretend that we do have it. So if I compute all the density to eps_0 .

Pretend that in fact I actually know eps_0 . Then I can compute every eps. If I know eps_0 I can compute that everyone else. I can compute the density so! I will get conditional maximum likelihood estimate. Conditional because I go t conditional model but also we stick with $\text{eps}_0 = 0$. This is not the same MA we have at the start, but we can compute solution. I transform huge complex model into a simpler model. It's not the model that we want, but will it be a good idea? We will see this later. If we think of doing this (DISEGNO di prima). When I got 1000 observation I will have $\text{eps}_0 + \theta^q 1000$. (it's exactly the starting process. Making false assumption I get easy solution. If MA(q) I will have to know the first q guys.

In ARMA(p, q) will be combination for the arguments.

We will also be interested in estimating θ .. but not σ^2 . θ is not interesting. σ is only not relevant for my estimation of FI: I can ignore all together. This way of ignoring parameter is called concentrating: we concentrate on the parameter we are interested in.

With a extra little assumption I manage to take this conditional in to something of Residual sum of square. This last page throws the natural conclusion. Even without the assumption of normality. So we don't need the assumption of normality to write this formula. If we see many software we actually call it Lisquare. Pseudo maximum likelihood if I have normally density but we don't. So pseudo maximum in: is not actually the maximum likelihood but it's similar. So this is what statistician will call it. It's important by appealing with this pretty Bernoulli problem and binomial formula and likelihood for binomial problem. In reality we actually know the distribution of the observations. So there's no chance to know if the guys are normally distributed or not. The normal distribution will be a big demand to impose to the data. If we are will to use this formula we use something called residual square.

PML when it is not normal will turn out to have the same proprieties that we would have if observations would be normal distributed.