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 referrer 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 **likehood**

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) = !n /(n-x)! x! p^x (1-p)^ nx → binomial.
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 int's not, the binomial is a function of X. Now on the other hand this is the function of Theta and we think that we know X so it's a different function. This is a function of Theta not X. This way of changing the function transform the prob in a new thing that is called likehood.

Where is the maximum theta?
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 likehood will take value for value from 0 to 1 and this approach is called maximum likehood.

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 likehood to happen.

What he don't show is that I can daw a picture and fin 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.
Ut is normally distributed, and we get the normal. It's Y – the mean. The product of all marginal is the join distribution.

The second one is a function in function of alpha beta and sigma squared. They look the same but they are not the same. Changes the observations.

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

The maximum of likehood. Instead of maximizing that function we take the logarithm of the likehood that is a monotonic function. Then I will call my estimate as I called before: Alpha, beta sigma are estimations of these parameter. We can solve it mathematically.
The derivative of u respect to alpha will be 1 and the derivative of beta is X. So we have twice the derivative of alpha and u.

I have alpha and beta unknown. I solve it and get estimation of alpha and beta. Moreover, using the fact that deviation for the mean is adapt to 0 so beta 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 vale of beta (as we didn't got the true probability of passing). As I change the observation I get a different observation for beta. So beta is itself a random variable with distribution that is a normal distribution.
This problem that we solve pointed out that the maximization of the likehood to minimizing the function on the right of the logarithm. So we have to minimize the sums of ut(alpha, beta) ^ 2. This gives me a way to estimate alpha and beta 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 ad 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 ut(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 likehood 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 rad(2 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 Yn. 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 likehood].
The square appear because the vector y-u' is multiplicated 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), theta and 0 everywhere else. I will factor element outside. How theta fits in the matrix of Omega. I know what the value of theta is.
If I'm interested in estimation the mean, mu 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 theta in bold not Beta]
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 theta. I have to do the Omega matrix. I take one potential value (0.5 for example) and I stick it to the likehood function. Then we take other value for theta. **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 theta NOT NICE. This is a 4x4 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 likehood 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 Y2. I could also use that the join density is the conditional density * times the marginal and fY2|Y1. Y1 is threated as we now it because is not random.

I can break the join density in the product of all conditional * the marginal.Yt-1 Is the only interesting. To turn this density in likehood we express for of Y.
The problem 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 likehood in the AR example before. Instead of Yt we called it ut. 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 likehood in every possible point of T. It does not work for a bit. The maximum of T (poi boh non ho capito).
Let's invent a new estimate and forget about the ln line and work on the second. What does it mean to do that?
The problem of the second line as a density. Everything given Y1. So really If I forget about the first line, I do a maximum likehood in a different design. Y1 is not random but it's given. This guy he proposes us is a different likehood with a different design. Y1 is not random. We called it **conditional maximum likehood.** (it's different form the maximum likehood).

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 likehood and take away a bit of Maximum Likehood 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 (Epst-1). If I conditional on Epst given Epst-1 gain 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 Yt-1. This is something that I observed. But I don't observe Epst-1. Maybe I can compute it? **How do I compute Epst-1? I could compute looking at his previous values.**
For example
I would like to use: Yt = eps t + theta epst-1. I want to get Epst-1.
Let's look at Yt-1 = eps t-1 + theta epst-2
Epst-1 = Yt-1 – Theta Eps t-2

The problem is that I trade Epst-1 with Epst-2 that I don't know.
So Yt = Epst + theta Yt-1 – theta^2 epst-2.
I still have epst-2 in way and it's not known. 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 Y1 and we get Eps0 that we don't have. But now let's pretend that we do have it. So if I compute all the density to eps0.
Pretend that in fact I actually know eps0. Then I can compute every eps. If I know eps0 I can compute everyone else. I can compute the density so! I will get conditional maximum likehood estimate. Conditional because I go t conditional model but also we stick with eps0 = 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 Eps0* theta*1000. Not 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 theta 1 .. but not in sigma^2. Theta1 is only on the right and what happened on the left of logarithm m is not interesting. Sigma is 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 likehood if I have normality density but we don't. So pseudo comes in: is not actually the maximum likehood 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 likehood 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 of 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.