Lezione 11 - 28/10/2019

venerdì 1 novembre 2019 19:54

Lezione 11 - 28/10/2019

Chapter 8: model selection

When we have the data we have to fit this guy (grafico). If i look at the correlogram i got an idea of the data.

Model selection

How do we choose the lags p,q in an ARMA(p,q) model?

- by looking at the sample autocorrelations and the sample partial autocorrelations, and trying to recognize the pattern of a model with given p,q.
- by using an automatic selection criterion (information criterion).

Simply looking at the correlogram it could be the starting point. The are mean 0 but they are not exactlly 0. This bar checking where is significant. So, the first step is to attach some probability to the correlogram. If the value is 0, this guys should be normalized.

Tests of "randomness"

If Y_t is i.i.d. (and has finite variance) then $\rho_1,\dots,\rho_k$ are all 0.
Then, the sample autocorrelations ($\hat\rho_j, \hat\rho_h$, $j \neq h, j \geq 1, h \geq 1$) are asymptotically independent and

$$\sqrt{T}\hat\rho_j \to_d N(0,1) \ (j \geq 1)$$

We can use this property to design two tests to check if the data are independently distributed.

"Test for randomness".
This test is so simple that it can be inspected visually, so the computers usually plots two error bars at $\pm 1.96/\sqrt{T}$ with the sample autocorrelation function.
(Notice: although it is called "test for randomness" by some computer softwares and some references, a more appropriate name would be "test for independent distribution").

Test if each has a normal distribution. Square root in the variance and divide it to the square root. And compare 1.96 / rad(T) and it's exactly what we're doing when we are testing.

"W" graph we can read as significant.

This test goes by the name of test of randomness. It's a bad name for a test, because regard if process is independent process. The nice thing is that is very easy to implement.

There is a variation of this test that is very interesting.

Portmanteau test
We can also test a group of k autocorrelations jointly: under the null,

$$T\sum_{j=1}^{k}\hat\rho_j^2 \to_d \chi^2_k$$

(this test may be of particular interest when we suspect a seasonal structure in the data: for example with quarterly data the first three autocorrelations may be zero, and then the fourth one may be non-zero). (The test may be sensitive to the choice of k on some occasions).

★ The test for randomness and the Portmanteau test can also be executed using the sample partial autocorrelations.

The tests for independent distribution and the Portmanteau test may provide preliminary information about the sample AC/PAC.

Examples

$T = 100$, $1.96/\sqrt{T} = 0.196$

	1	2	3	4	5	6	7	8	9	10	11	12
$\hat\rho_j$	-0.041	0.005	0.150	0.116	-0.027	0.048	0.072	0.020	0.155	-0.052	-0.090	0.209
$\hat\alpha_j^{(j)}$	-0.041	0.003	0.150	0.132	-0.017	0.021	0.040	0.018	0.158	-0.064	-0.125	0.164

Portmanteau (12) = 12.47 (c.v. 21.02)

	1	2	3	4	5	6	7	8	9	10	11	12
$\hat\rho_j$	0.631	0.478	0.448	0.365	0.257	0.251	0.240	0.223	0.229	0.133	0.103	0.194
$\hat\alpha_j^{(j)}$	0.631	0.133	0.173	0.004	-0.062	0.074	0.042	0.050	0.057	-0.143	-0.001	0.176

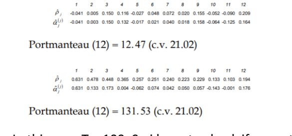
Portmanteau (12) = 131.53 (c.v. 21.02)

In this case T = 100. So i have to check if my autocorrelation is below or above. I have 24 hits and one of the is insignificant. I'm doing test and i got critical value of 0.196 i will have 5 % change that something going beyond the critical value. So 5% accepting something that i don't want. If i run the 24 times, i aspect there is one hit. 1/24 is not a surprise. So i need a bit of flexibility reading this computer output of the autocorrelation. There is an easier way.

This guys s not certainly and independent process (Example 12.47)

There is a way to aggregate all these numbers in just one statistic. Pretending all guys are from standard normal. CHI square 1,2,3,4 to 12 summing all the values. I can consider them together I aggregate them, i squared them so i eliminate the sign and then i sum them to get a chi square 1,2 to 12. SO instead of looking them individually i can look at the sum of the square.

This test is called Portmanteau test and it's a test useful for a lot of situations.



How do i choose the right value?

This is a convenient model to look at. The idea if i want to find out p and q i have to think a little bit more. How do you normally choose between two model? Let's make an example.

Let's start with and example:

AR(1) vs AR(2)

AR(2) is $Y_t = \varphi_1 Y_{t-1} + \varphi_2 Y_{t-2} + \varepsilon_t$

AR(1) is $Y_t = \varphi_1 Y_{t-1} + \varepsilon_t$

ESTIMATES $\to \hat\varphi_1$ $\to \hat\varphi_1, \hat\varphi_2$

Even when fi2 should be 0, the estimate wouldn't be zero because the estimate is a random variable with a normal distribution. It will be very close to 0.

AR(2)

$$\min_{\varphi_1,\varphi_2} \underbrace{\sum (Y_t - \varphi_1 Y_{t-1} - \varphi_2 Y_{t-2})^2}_{\varepsilon_t(\varphi_1,\varphi_2)}$$

AR(1)

$$\min_{\varphi_1} \underbrace{\sum (Y_t - \varphi_1 Y_{t-1})^2}_{\varepsilon_t(\varphi_1)}$$

How we will choose FI1 and F2? In the way to get the min of sums squares. We can do it numerically and focus on the fat that is a function of FI1 and F2 with a three-dimensional parameter. We could just think all possible value for F1 and F2. This procedure many times. We choose our estimate as the pair that give us the low value of this pairs. Than after we got all best value of FI1 we got the minimum.

FI2 estimates that is -0.1, is better than having 0 as estimates. Lower Mssquare with 0.1 instead of 0. With probability 1 is lower than the best possible AR(1). The answer is bigger aumentando il numero di regressioni.

I estimate the bigger model and i test in this example: i estimate the AR(2) and i estimate FI2 == 0. If the estimation of FI2 is not significant i will go with the AR(1) model. This works well on regression. We'll be able to compare them, but how versus MA(2)??

MA(2) Yt = Eps + theta1 epst-1 + theta2 Epst-2

The argument before is i can test and check if parameter is 0. It works when i test AR(2) with AR(1). But can i compare AR(1) to MA(2)? There is no way to restrict the parameter of AR(1) and get the model of MA(2). SO this model are not comparable. So this works when the model are comparable. This is called nasty. So ARMA models cannot be compare is they are not comparable.

There is an information of the like hood. I cannot check the like hood of MA and AR. Like hood would be higher in the case of AR(2) between AR(1). We will compare the maximize like hood ad we will think that the bigger model has an advantage: lower Min of sums square and bigger like hood. And that's is presented on this Slide:

Model Selection - Information criteria

an automatic way to select q,p.
The idea: use "maximum likelihood" to choose p,q.
The problem: if you compare an ARMA(p,q) with an ARMA$(p+1,q)$, the ARMA(p,q) has always less likelihood.
This is because the estimate from the ARMA(p,q) model maximises the likelihood with the constraint that $\hat\phi_{p+1} = 0$, while the ARMA$(p+1,q)$ does not impose that constraint, so the ARMA$(p+1,q)$ has higher maximum likelihood unless $\hat\phi_{p+1} = 0$ exactly (which is an event with probability zero in finite sample even when the true $\phi_{p;0} = 0$ actually) (Notice analogy with regression here: when you increase the number of regressors, the R^2 does not decrease, and in general increases, even when the regressors are irrelevant).

ARMA(p,1) and ARMA(p+1, q) will have more like hood

The solution: add a penalty which increases with p and q.

$$IC = -2\mathcal{L}(\hat{\boldsymbol\beta}) + penalty$$

$$penalty : \begin{cases} 2(p+q) & \text{Akaike IC} \\ (\ln T)(p+q) & \text{Bayes IC} \end{cases}$$

BIC: consistent estimation of p, q.
AIC: inconsistent estimation of p, q (may select larger than correct p, q in large samples).
Both BIC and AIC may select smaller then correct p, q in finite samples (this however is not necessarily a bad thing: it may result, in small samples, in smaller forecast MSE).

An alternative approach: of course, we can also compare an ARMA(p,q) with an ARMA$(p+1,q)$, or with an ARMA$(p,q+1)$, using a likelihood ratio test. The criterion is then adding lags as long as the likelihood ratio test statistic is above a user-chosen critical value (for example, 5% significance would have c.v. 3.84).

We have two penalties possibilities. Akaike information criteria and Bayes information criteria.

If we go to the example Z and verify this procedure:

Information Criteria example

Example: automatic lag selection for Z

	p+q	l-lik(1)	AIC	BIC	LR
iid	0	-172.99	345.98	345.98	
MA(1)	1	-146.17	294.35	301.56	53.63
AR(1)	1	-140.28	282.57	289.78	65.41
ARMA(1,1)	2	-136.15	276.31	290.73	20.04(2), 8.26(3)
MA(2)	2	-131.15	266.30	280.72	30.05
AR(2)	2	-135.40	274.79	289.21	9.77
MA(3)	3	-130.96	267.91	289.55	0.39
ARMA(1,2)	3	-130.64	267.29	288.92	11.02(4), 1.01(5)
ARMA(2,1)	3	-135.37	276.74	298.37	1.57(6), 0.05(7)
AR(3)	3	-134.39	274.78	296.41	2.01

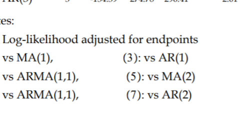
Notes:
(1): Log-likelihood adjusted for endpoints
(2): vs MA(1), (3): vs AR(1)
(4): vs ARMA(1,1), (5): vs MA(2)
(6): vs ARMA(1,1), (7): vs AR(2)

AR(1): as i increase like hood increasing the penalty.

I can do it for all the model and parameters. This procedure makes me compute 10 model instead of computing every value. Lowest value is 280 but now the best for the like hood.

MA(1) vs MA(2) $H_0 : \{\theta_3 = 0\}$

$Y_t = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \theta_3 \varepsilon_{t-3}$



When we look at these SLIDE with numerical examples we wouldn't normally do that. We understand how to compute AIC and BIC. There is one problem: if i have both AIC and BIC gave me same number, but if i have different model and numbers? Every choice will be good, but we will never do it in practise. We will have a process that will calculate this criteria. If i have something that given me wo different answer how do i choose? I can choose thanks to this information criteria. The BIC five a consistent estimation for p and q which means that if i have a very large sample the BIC will return the right value of p and q but the AIC will not. If large samples will select a model that is bigger than it needs to be. I will always choose for BIC then. But if i ho for AIC is equally fine but BIC works better.

We find two ways to find p and q. But what is the value of p and q that i really want to use? I want to use the p and q to simplify the model to get the best forecast. And something we're better of simplify the model.

EXAMPLE

1. Adding non-necessary parameters results in larger variation of the estimates, i.e. the estimates (and the forecasts) are not precise.

We can see this easily in the AR(1) example:

$$Y_{t+1} = \phi_1 Y_t + \varepsilon_{t+1} \text{ (model)}$$

Suppose that we fitted the AR(2),

$$\hat{Y}_{t+1|t,\dots} = \hat\phi_1 Y_t + \hat\phi_2 Y_{t-1} \text{ (forecast)}$$

then

$$Y_{t+1} - \hat{Y}_{t+1|t,\dots} = \left(\phi_1 - \hat\phi_1\right)Y_t + \left(-\hat\phi_2\right)Y_{t-1} + \varepsilon_{t+1}$$
(forecast error)

Fitting the AR(2) instead of the AR(1) increases the variance of $\left(\phi_1 - \hat\phi_1\right)$ and adds the variance of $\hat\phi_2$ to the forecast error. This means that the forecast MSE is larger when the AR(2) instead of the AR(1) is used.

I will have the variance of the errors. The more parameter we estimates the more variance we stick in the model. So, estimates a model that is bigger will increase variance for no reason.

Adding non-necessary parameters results in larger variation of the estimates, ie the estimates (and the forecasts) are not precise.
- Example 1. AR(1).

The series

was generated as AR(1) with $\phi = 0.75$.
★ If we pretend not to know the true model, and that we are uncertain between an AR(1) and an AR(2), we estimate ϕ with both models.
Call $\hat\phi_{1AR(1)}$ the estimate of ϕ when the AR(1) is assumed, and $\hat\phi_{1AR(2)}, \hat\phi_{2AR(2)}$ the estimates of ϕ_1 and ϕ_2 when the AR(2) is assumed. We found

$$\hat\phi_{1AR(1)} = 0.747, \quad \hat\phi_{1AR(2)} = 0.729$$

so in this particular example $\hat\phi_{1AR(1)}$ got closer to ϕ, so the AR(1) worked better.

I estimate model with AR(1) and i get 0.747 and i will make a forecast. If i have AR(2) i got 0.729 and is not good as the AR(1) in this particular example. Run the forecast and the forecast for AR(1) is -0.93 and AR(2) -0.94 and the forecast for the smaller is better. How many parameter i want to estimates? Not so many because the more i have the more is the variance! Don't go past the MA(2). But this is just one experiment. If i do it 1000 times

★ If we forecast Y_{T+1},
AR(1) $\hat{Y}_{T+1|T,\dots} = \hat\phi_{1AR(1)} Y_T$
AR(2) $\hat{Y}_{T+1|T,\dots} = \hat\phi_{1AR(2)} Y_T + \hat\phi_{2AR(2)} Y_{T-1}$

In our example,
$Y_{T+1} = -0.34$
AR(1) $\hat{Y}_{T+1|T,\dots} = -0.93$
AR(2) $\hat{Y}_{T+1|T,\dots} = -0.94$

so in this particular example the AR(1) gave the best forecast.

- Example 2. 1000s AR(1), an experiment.
Take 1000 different (random) similar series:
★ the estimate $\hat\phi_{1AR(1)}$ results to be closer to 0.75 than $\hat\phi_{1AR(2)}$ in 58.7% of the cases;
★ the standard error of the estimated values $\hat\phi_{1AR(1)}$ is 0.072, the standard error of the estimated values $\hat\phi_{1AR(2)}$ is 0.101.
★ the forecast $\hat{Y}_{T+1|T,\dots}$ from AR(1) results closer to Y_{T+1} than from AR(2) in 54% of the cases;
★ the standard error of the forecast error $Y_{T+1} - \hat{Y}_{T+1|T,\dots}$ from AR(1) is 0.968, the standard error of the forecast error from AR(2) is 0.977.

Let's look at this example.

Sometimes, using a smaller model may even give more precise forecasts than the correct model.
- Example 3. AR(2), an experiment.
Suppose now that we have 1000 series from
$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \varepsilon_t$, $t = 1,\dots T, T+1$
with $\phi_1 = 0.65$, $\phi_2 = 0.1$ and we consider again:
using $t = 1,\dots T$ to estimate ϕ_1,ϕ_2, in AR(2) and then forecast Y_{T+1};
using $t = 1,\dots T$ to estimate ϕ_1 in AR(1) and then forecast Y_{T+1}.
★ when $T = 100$, the forecast $\hat{Y}_{T+1|T,\dots}$ from AR(1) results closer to Y_{T+1} than from AR(2) in 50% of the cases;
★ the standard error of the forecast error $Y_{T+1} - \hat{Y}_{T+1|T,\dots}$ from AR(1) is 0.996, the standard error of the forecast error from AR(2) is 0.997.

Of course, this depends on the fact that T is small and ϕ_2 is small: both things make estimating ϕ_1 and ϕ_2 in the AR(2) not precise, and therefore the forecast is better with an AR(1). With larger T and larger ϕ_2 the result would be better for the AR(2) model.

The AR(1) is too small, so it bound to not to be consistent becuase impose thewrong value for FI2. But if we run the race the AR(1) has a better forcast. Why is that? Aldthoguht FI2 is 0.1 and for AR(1) we stick 0. SO the price of having FI2 = 0 is lesser than the incresing the variance during the estimations.

Another example

FI = -0.55 and Theta = 0.45

- Example 4. ARMA(1,1), an experiment.
Suppose now that we have 1000 series from
$Y_t = \phi Y_{t-1} + \varepsilon_t + \theta\varepsilon_{t-1}$, $t = 1,\dots T, T+1$
with $\phi = -0.55$, $\theta = 0.45$ and we consider:
using $t = 1,\dots T$ to estimate ϕ, θ, in ARMA(1,1) and then forecast Y_{T+1};
using $t = 1,\dots T$ forecast Y_{T+1} assumung that Y_t is an independent process (the rationale for this is that $\phi = -0.55$, $\theta = 0.45$ is very close to $\phi = -0.5$, $\theta = 0.5$, in which case we would have a common factor so actually Y_t would be an independent process).
★ when $T = 100$, the forecast $\hat{Y}_{T+1|T,\dots}$ from iid results closer to Y_{T+1} than from ARMA(1,1) in 51.3% of the cases;
★ the standard error of the forecast error $Y_{T+1} - \hat{Y}_{T+1|T,\dots}$ from iid is 1.010, the standard error of the forecast error from ARMA(1,1) is 1.028.

Of course, this depends on the fact that T is small and $-\phi$ and θ are close to each other. With larger T the estimates would be more precise and the result would be better for the ARMA(1,1) model.

The simplify win in sense of having a better forecast and a better version of the forecast.

$Y_t = 0{,}55\,Y_{t-1} + \varepsilon_t + 0{,}45\,\varepsilon_{t-1}$

$(1 + 0{,}55L) = (1 + 0{,}45L)\varepsilon_t \rightarrow$ THEY ARE NOT THE SAME BUT SIMILAR

This example show that not only getting a model small, but also a model that is really smaller than the real p and q. Smaller model tends to perform better than bigger ones. So when i go back and i comparing the Bayes tends to get a smaller example. This means that it will going to get the better forecast.

We call this model Parsimonious model, because we have to be parsimonious.

Parsimonious modelling

Large econometrics models tend to do badly in terms of forecasting, and are outperfomed by small ARMA models (Box & Jenkins).
Even in ARMA models, increasing the number of parameters reduces the precision of with which each parameter is estimated: this may worsen the MSE. This is because when the parameters are estimated, their variance contributed to the variance of the forecast. Adding extra parameters may then help to reduce or eliminate the forecast bias, but the gain in terms of reduction *bias*^{2} is outweighted by the loss in increased *variance* of the forecast.
Should balance the number of estimated parameters and the number of observations.
Sometimes, Information Criteria have been advocated also to select more parsimonious models.

Last thing: suppose that i look at series Z and i look it's an MA(2) and i got estimate and select model using criteria and so i estimate the model and i get the MA(2). This could be the end of the story but we want the model to be small. The estimate of Eps (model validation).

Model validation

We just estimated $\hat{\boldsymbol\beta}$ for an ARMA(p,q). We can then compute the residuals

$$\varepsilon_t(\hat{\boldsymbol\beta}) = Y_t - \hat{c} - \hat\phi_1 Y_{t-1} - \dots - \hat\phi_p Y_{t-p} - \hat\theta_1\varepsilon_{t-1}(\hat{\boldsymbol\beta}) - \dots - \hat\theta_q\varepsilon_{t-q}(\hat{\boldsymbol\beta})$$

(initialising the sequence setting $\varepsilon_p = \varepsilon_{p-1} = \dots = \varepsilon_{p-q+1} = 0$ as usual): if the data are really ARMA(p,q), the residuals $\varepsilon_t(\hat{\boldsymbol\beta})$ should approximate well the true ε_t.
Introduce for the residuals the abbreviation

$$\hat\varepsilon_t = \varepsilon_t(\hat{\boldsymbol\beta})$$

and consider the sample autocorrelation of the residuals

$$r_j = \frac{\frac{1}{T}\sum_{t=j+1}^{T}\hat\varepsilon_t\hat\varepsilon_{t-j}}{\frac{1}{T}\sum_{t=1}^{T}\hat\varepsilon_t^2},$$

then the Portmanteau statistic for the sample autocorrelation has limit distribution

$$T\sum_{j=1}^{k} r_j^2 \to_d \chi^2_{k-(p+q)}.$$

BIC and AIC is an estimation but there is not optimality. The test is takin this Eps and throw them in the Portmanteau statistic. The final step is taking the residuals, calculate autocorrelation and then compute the Portmanteau statistic.

EXAMPLE

Model validation

Correlograms of the residuals when we fitted either a MA(1) or a MA(2) to Z.

For example, when $k = 3$ lags are selected, we can compute the Portmanteau statistics as
$MA(1)$ residuals: (asy. χ^2_2 under no autocorrelation)
$100 \times (0.285^2 + 0.321^2 + 0.110^2) = 19.637$
$MA(2)$ residuals: (asy. χ^2_1 under no autocorrelation)
$100 \times (0.039^2 + 0.027^2 + 0.041^2) = 0.3931$

Under the assumption of no residual autocorrelation, the Portmanteau statistic is asymptotically $\chi^2_{k-(p+q)}$ distributed.
In this example, this distribution is χ^2_2 when the MA(1) is fitted, and χ^2_1 when the MA(2) is fitted.
The 5% critical values are 5.99 for the χ^2_2 and 3.84 for the χ^2_1.
Thus, the assumption that the residuals are not autocorrelated when the MA(1) is fitted is rejected. On the other hand, when the MA(2) is fitted, the assumption is not rejected.

This backward test seals the estimation procedure. We check with the test at the end.