Time Series Modelling and Forecasting with Applications in R

Prof. Sudeep Bapat

Shailesh J. Mehta School of Management

Indian Institute of Technology Bombay

Week 05

Lecture 22: Measuring Forecast Accuracy

Hello all, welcome to this new lecture in the course on time series modeling and forecasting using R. Now, up until the first lecture this week, we have kind of started with exploring more about forecasting, right. So, before that, we have tried to analyze some real data, then tried to model the real data using some basic time series models, then tried to compare the model fit using some information criteria, and so on and so forth. And then lastly, we have kind of applied some diagnostic checking to ensure that all the assumptions about the model are being met or not. And once you are through with all these stages, then the ultimate goal is forecasting.

Now, in the last lecture, we saw a particular example where we tried to forecast, let us say, one-step-ahead forecast or two-step-ahead forecast or, in general, L-step-ahead forecast of a particular ARMA model. Now, one last thing we will do with the ARMA model again, and then we will see what would be covered next. Now again, if you vaguely remember from the last lecture, we tried to forecast, in general, the L-step-ahead forecast for a particular ARMA model with orders P and Q. And then on top of that, we found out the forecast error also, and then we talked about, let us say, the expected value of the forecast error and, let us say, the variance of the forecast error, and so on and so forth. So here, interestingly, we have two properties about any forecasts.

So one can actually note that As n tends to infinity, the actual value of the forecast sort of tends to the overall mean of the process, which is mu. So essentially, if you want to visualize this idea, right? So if the idea is to forecast a particular ARMA model, then eventually what would happen is that even though the series looks like that, right? Then let us say the forecast is something like that, but eventually, it will kind of converge to the mean, right?

So, let us say the mean is somewhere here, and eventually, all the forecasts would be exactly equal to or really close to the overall mean of the series. So, this is a kind of disadvantageous tendency of any ARMA structure, right? In the longer term, it would not give you better and better forecasts, but the idea is that the forecasts would only converge to the overall mean. At the same time, as n goes to infinity, the variance of the forecast error, right? Now, again, remember what exactly the forecast error is. So, the forecast error is nothing but the actual value minus the forecasted value.

For example, something like E N L is the actual value. So, Y N plus L minus the forecasted value, something like Y N L and then hat. And then we can talk about its expectation and its variance. So, by the way, the expected value of the forecast error happens to be 0 under an ARMA structure. We have seen that in the last lecture.

But what happens to the variance of the forecast error? Now again, as n tends to infinity, the variance of the forecast error also converges to the overall variance, which is gamma 0. So, gamma 0 is what from notation? So, gamma 0 is nothing but the variance of the actual time series process, is it not? So, gamma 0 is nothing but the variance of yt.

So, in the long run, the forecast would kind of slowly converge to the actual mean, while the variance of the forecast errors would converge to the actual variance of the time series process. Okay. Thus, this may be a slight drawback of an ARMA model or ARIMA model and so on. Right. So, thus ARMA or ARIMA forecasting is useful only for short-term forecasts.

So, I think this point is really important that if you want to forecast something in the extreme future or let us say down the line, then probably using an ARMA model or ARIMA model would not be a good fit. Right. Because ARMA forecasting or ARIMA forecasting is useful only in short-term forecasts, right? So, then what can we have as some other better technique? So, we will slowly progress this week by analyzing all the different ideas about forecasting time series data, okay?

Alright. So, now the immediate next question is, can you create some prediction intervals? So, once you forecast something or once you predict something, rather than giving me a point estimator, can you sort of give me a confidence interval around the unknown parameter, right? And again, the answer is yes. So, we can actually do it.

$$\hat{Y}_n(l) \pm 1.96\sqrt{Var(e_n(l))}$$

$$= \hat{Y}_n(l) \pm 1.96\sigma_a \sqrt{\sum_{i=0}^{l-1} \psi_i^2}$$

So, a 95% prediction interval for Y n plus L. Now, again, remember Y n plus L is the value that we want to forecast, right? So, Y n plus L is nothing but the L step ahead value which needs to be forecasted, okay. So, can we sort of propose a prediction interval around Y n plus L? So, again, the answer is yes that a 95 percent such P I. So, in short, we will call this prediction interval as P I. So, 95 percent P I is given by nothing but the actual forecasted value which plus or minus 1.96. Now, again, 1.96 is a routine value connected to a 95 percent interval which is nothing but the z-score, right, and into the standard error of the forecast error or nothing but the variance of the enl and then under root. So, this is nothing but the standard error, okay? This is nothing but the standard error.

$$\hat{Y}_n(l) \pm 1.96\sigma_a$$

So, essentially, what do you have? So, you have the actual forecast plus or minus 1.96, right, and into the variance of the forecasted error and then under root of that, right. So, in particular, I can actually write down in this equation form. So, this is the forecasted error plus or minus 1.96. Now, this entity, so under root variance of the forecast error takes this form.

So, sigma a and then under root summation psi i square i going from 0 to L minus 1, and this we saw in the last lecture, is it not? So, the variance of the forecast error is nothing but given by this form. So, the only thing you have to do is you have to take an under root of that which would denote the standard error of the forecast error. So, here the entire game is about capturing the forecast error that you are making and then converting that to either a variance form or a standard error of the forecast error and then multiply that value with the z-score. So, 1.96 and then plus and minus the actual forecasted value.

Thus, if you again plug L to be 1, that would give me a one-step-ahead forecast. Thus, a prediction interval for a one-step-ahead forecast would be nothing but Yn L hat plus or minus 1.96 into sigma A only. Because here, remember that I would go from 0 to 0 because L would be 1, right? So, there will just be a single value, which is psi 0 square, by the way. And then, psi 0 square I can trivially take to be as 1.

So, essentially, the prediction interval for a one-step-ahead forecast would be nothing but this guy. And then, similarly, I can create prediction intervals for, let us say, two-step-ahead forecasts or three-step-ahead forecasts or, in general, L-step-ahead forecasts. All right. Now, I think this slide kind of presses more on the fact as to why one needs some long realizations of the actual time series process. So, firstly, long realization means what?

So, one should be able to collect a lot of observations and then not just a handful. So, long realization means that the actual sample data is really long. So, it contains lots and lots of observations. So, a few advantages of obtaining such samples, having a lot of observations or having a long realization, are all these. So, let us say, estimate correlation structure, let us say, either ACF or PACF.

So, estimating such structures, let us say ACF and PACF functions, is easier, and one can actually get accurate standard errors. So, of course, if the sample size is large, if you are collecting more observations, estimating the exact correlation structure using either the ACF form or the PACF form becomes easier and more accurate. So, the standard errors of the estimates are also accurate. Better prediction intervals of realization are large. Now, again, all these are general kinds of statements, right.

So, if you are collecting more observations or if n is large, then the prediction intervals are also better. They are more accurate. Fewer estimation problems when it comes to, let us say, MLE or when it involves a likelihood function, right. So, when the sample size enlarges, the likelihood function is better behaved, right. So, you can sort of model the likelihood function or get the MLE, which is more accurate, or the likelihood function is better behaved.

So, one actually has fewer estimation problems in that aspect. Yeah, it is possible to check forecasts by withholding recent data. Now, in the era of machine learning, data science, and all these things, what people do is they keep some data to train the model. So, this is called a training data set, and then they keep a handful of observations for testing the model, right. Now, again, if you are having only, let us say, 30 observations or 40 observations to start with, then splitting 30 observations or 40 observations into a training set and a testing set is not feasible. But, of course, if you have lots and lots of observations, let us say hundreds or thousands and so on, then one can appropriately split the data into a training set and a testing set.

And again, just to summarize what exactly these two mean is that whatever model you are fitting, let us say ARMA, ARIMA, SARIMA, etc., the models have to be fitted on the training set only. And the testing set has to be used to check if the model is accurate or not by obtaining the errors or obtaining the deviations from the actual value in the test set and the fitted values or the forecasted values. So, this is the idea about splitting any data set into a training set and a testing set. So, again coming back to the same point, just to summarize, is that if you are having long realizations or lots of observations then It is possible to check the forecast by withholding some recent data and then keeping some recent data in the testing set and keeping all the prior data in the training set.

And lastly, if you have more observations, one can actually check the model stability by dividing the data and analyzing both sides. So, again, these are some points as to if you have long realizations of any time series data, then what might be the advantages of that. Now the next idea is the advantages of parsimonious models. Now you have to understand what you mean by parsimonious models. So a parsimonious model means a model having fewer parameters but still accurate.

I will give you an example. So again, think of a situation where you are trying to fit A real data set using two possible models. So the first one is AR2 and then the second one is let us say AR8. Right.

And now let us say you are comparing these two models using some AIC structure or BIC information, whatever. And for some reason, let us say the AIC values come out to be, let us say, 100.5 in this case, and then the AIC value for the AR8 model comes out to be, let us say, 99.8, alright. Again, just to summarize, for a particular practical data, you are trying to fit two models. AR2, let us say, and then AR8, let us say, right. And then, what exactly are these numerical values?

These are AIC values for fitting both these models. So, the AIC value for fitting the AR2 model is 100.5, and then the AIC value for fitting the AR8 model is 99.8, right. Now, of course, if you close your eyes and then pick the better model only related to the AIC criteria, then which model would it be? So, of course, it will be AR8 because the AIC is lower here. But does that mean that you have done a very good job?

Probably not, because AR2 is a really complicated model. So, what is the order of that? So, the order is 8, firstly. So, how many parameters would it have? It would have 8 parameters as compared to only 2 parameters here.

And on top of that, if you see the difference in the AIC, the difference in AIC is very, very negligible. So, 99.8 and then 100.5. So, in this case, a parsimonious model would be nothing but AR2 because AR2 is doing almost an equal job, I will say, because the information criteria is not very far apart from 99.8. And at the same time, it only contains just two unknown parameters. And remember one thing: if you have more unknown parameters in the model, then estimating all the unknown parameters will result in some errors associated with the estimation and so on, right?

So, the error kind of propagates if you have more parameters, right? So, this is the idea of a parsimonious model. So, a parsimonious model means that it should contain a lesser number of parameters, but it should be almost as accurate as a non-parsimonious model. So, a few advantages of using a parsimonious model are fewer numerical problems in estimation. So, we have seen that, right, because the AR2 model is a simplistic model.

So, there would be fewer problems when it comes to estimation. Easier to understand the model. Now, again, if you show the model structure of these two models to some layman, right? Now, he or she might be afraid by looking at the structure of AR8, right? Because it involves all those eight parameters and so on and so forth.

But at the same time, the AR2 model has a very easy structure to visualize, okay? Third, with fewer parameters, the forecast would be less sensitive to deviations between parameters and estimates, right? Again, AR2 only contains phi 1 and phi 2, which need to be estimated, right. So, when it comes to the overall forecast error, right. The overall forecast error would be less sensitive to deviations between, let us say, phi 1 and phi 1 hat and phi 2 and phi 2 hat, as compared to all the 8 parameters and their estimates for AR8.

Next, the model may be applied more generally to similar processes. Since you are dealing with a simpler version or a simpler process, you can actually extend the modeling aspect of that simpler model to other similar processes also. Next is rapid real-time computations for control or other actions. So, since the model itself is easy to deal with, then computationally it also becomes slightly faster. And lastly, having a parsimonious model is less important if the realization is large.

So, the last point is, in fact, telling you that if you have a long realization, again, what do you mean by a long realization? You have lots and lots of observations. Then having a parsimonious model is less important because even if you have more parameters in the model, for example, let us say AR8, but at the same time, the realization is really long, then they kind of balance the effects of each other, right? So, think of a situation where

you are collecting lots and lots of observations. So, probably fitting AR8, if the AIC criteria suggests that, should be taken further for forecasting and whatever, as compared to AR8.

So, hopefully, these points were clear, which were seen in the last slide. Now, what we will do is, let us say, once you forecast a couple of models, right? So, let us say AR2 and AR8, then how do you actually measure the forecast accuracy? And then, using all these measures that you see in front of you of forecast accuracy, can you then compare between the forecasts? And then, can you say which model might be better and so on and so forth? Again, the answer is yes. So, firstly, we will have a quick review of what all measures one can use when it comes to quantifying the forecast accuracy. And again, just for a second, if you pause the video and then see in this slide, in this table here, the entire list of probable measures contains a single common term, which is error, right? So, mean squared error, mean error, mean percentage error, etc.

So, the idea of measuring the forecast accuracy has to revolve around measuring the errors, right? Now, which error again? So, error is nothing but the forecast error, which is nothing but ENL, okay? So, again, just to summarize, ENL is nothing but the actual value minus the forecasted value. Right? So, if the error is really small in some context, let us say mean squared error or mean error, mean percentage error, etc., then we can actually safely say that the forecast is accurate, right? And this could be a criterion to compare different models also, isn't it? I will give you an example. So, let us say again, you are dealing with practical data. You want to forecast down the line using two probable models. So, the first one is, let us say, AR2, as taken before, and then the second one is AR8, right?

Again, just to summarize, you have practical data. You are fitting these two models on the dataset, let us say AR2 and AR8, and then obviously you estimate the parameters, you do all the sanity checks as diagnostic checks, and so on and so forth. Then you forecast. Now, our task is to find out which forecasts are more accurate. Then, what we can do is, we can find out, let us say, the MSE of the forecast error for this model. We can find out the MSE of the forecast error of this model. And whichever MSE is least or whichever MSE is lower, we will kind of go with that model. So, in a way, measuring forecast accuracy is not to tell you how accurate the forecasts are.

Obviously, that is one of the reasons. The other advantage is that you can actually compare between different models also. All right. Now, again in front of you, if you go

back to this table here, you will see lots of different sorts of measures. So, the first one is mean squared error.

The abbreviation is MSE. And then, I think the last column tells you a brief description about the technique or about the measure. So, what exactly do you mean by MAC? So, MAC is nothing but the average of squared errors over the sample period. Next one is mean error.

The abbreviation is ME. So, what is mean error? So, it is the average percentage points by which the forecast differs from the outcomes. So again, if you notice, each one of these has some subtle difference compared to the previous one. Right.

Now, the next one is mean percentage error. So, in short, we can write it down as MPE. So, MPE means the average of the percentage errors by which the forecast differs from the outcomes. Next one is mean absolute error. So, mean absolute error abbreviation is MAE.

So, the average of absolute percentage points by which the forecast differs from the outcome. So, on one hand, we started with squared errors, measuring squared errors. And now here, when it comes to MAE or mean absolute error, we are trying to measure the absolute error rather than squared error. And lastly, mean absolute percentage error or MAPE, right, which gives you the average of absolute percentage amount by which the forecast differs from the outcomes, right. So, in some of these techniques, you will see that there is a percentage term added, and in some of the other techniques, the percentage term is not there, right.

So, these are some subtle differences. Now, in the subsequent slides, we will try to review the formulas for all these five techniques, right. So, we will start with MAC, then mean error, then mean percentage error, mean absolute error, and then lastly MAPE, right. All right. Now, again, we have to kind of stick to this fact or this assumption throughout the next set of slides.

So, let AT denote the actual value. And let FT denote the forecast value, right? So, FT is the forecasted value, and AT is the actual value. Now, again, I should not have used the AT notation here because, remember, AT denotes the error, right? In our case, but nevertheless.

So, let us say AT. So, A stands for actual, and F stands for forecast, okay? So, now, how exactly can you write down the mean squared error? So, this is the formula for mean

squared error. So, what you do essentially is you take the actual value minus the forecasted value, you take the square of that, and then you take a summation divided by n, okay.

$$MSE = \frac{1}{n}\sum_{t=1}^{n}(a_t - f_t)^2$$

So, 1 by n summation t going from 1 to n a t minus f t and then whole square, right. Now, again, how do you visualize all these errors again? So, let us say I will give you a small example. So, let us say you have some practical data that you have collected, right? This is the partition, and after that, the task is to forecast, right?

Now, let us say you forecast it for the next 12 months. So, there will be F1 value, F2 value, F3 value, and so on up to F12. So, consider a very simple situation. Let us say the idea is to forecast next year's temperatures, monthly temperatures, based on some historical data. So, up to this point, you have collected the monthly temperatures for, let us say, the past 5 years or something like that, and now the idea is to forecast the monthly temperatures for the next year.

So, F1, F2, F3, F4 up to F12 would be the forecasted temperatures on a monthly basis, and let us say the actual values for the temperatures are these. So, A1, A2, A3 up to A12. So, is the idea clear? So, the AIs are the actual values, and the FIs are the forecasted values. And then the idea is to find the error between the actual value and the forecasted value using some appropriate measure.

So now, for example, MSE could be one such measure where you simply have this simple-looking formula. Or the second one is the mean percentage error. Now, when it comes to percentage, you have to quantify the error in terms of some percentage. So, something like this. So, summation and then 80 minus FT divided by 80 into 100.

$$MPE = \frac{1}{n}\sum_{t=1}^{n}\frac{(a_t - f_t)}{a_t} \times 100$$

So, this would give you some percentage. And then again, summation and 1 by N. And so on. So, this is the idea behind all the different measures or all the different errors. Then the next one is mean absolute error. Now, this is slightly different, obviously.

$$MAE = \frac{1}{n}\sum_{t=1}^{n}|(a_t - f_t)|$$

So, mean absolute error would be 1 by n summation then the absolute value of a t minus f t. So, rather than a square term here, I have an absolute value, right? Then what do you mean by mean absolute percentage error or MAPE? So, I think MAPE is a kind of very widely applied measure. So, MAPE is what? So, 1 by n summation absolute value of 80 minus FT divided by 80 into 100.

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\frac{|(a_t - f_t)|}{a_t} \times 100$$

Now, if you notice one very easy thing, then the mean absolute error and the mean absolute percentage error, the only difference is you are kind of converting the mean absolute error in terms of percentage here. Similarly, if you go back, then the mean squared error and the mean percentage error is what? So, you are kind of converting the mean squared error to some sort of a percentage here. So, this is the idea, basically. So, mean absolute error, then mean absolute percentage error, and so on.

Now, lastly, we will discuss the idea of a naive forecast. So, what exactly does the idea of a naive forecast mean? So, firstly, what exactly does a naive forecast mean? So, a naive forecast is nothing but an estimation technique in which the last period's actuals are used as this period's forecast. Now, what do you mean by that?

So, let us say again you are collecting observations on a monthly basis, let us say temperature data, and you have collected lots of observations. So, y1, y2, y3 up to yn. So, now let us say yn would be the last month's temperature because yn is the most recent data you have collected. So, let us say yn is the last month's temperature. So, this is the most recent data.

Now, if I want to propose a very naive forecast, what I will do is I will take the one-step-ahead forecast to be the value Yn itself. So, this is the idea of a naive forecast. Because what has happened is in the last month, I know that the temperature was Yn. Let us say 35 degrees on average. So, a naive way of guessing, a naive way of forecasting the temperature of this month would be nothing but using the same exact value which was the temperature in the last month.

Now, obviously, this is not a very good idea because the idea of forecasting itself is kind of defeated. But then, the idea of a naive forecast is when the last period's actuals or the last period's values are used as this period's forecast without adjusting them or without attempting to establish causal factors, nothing like that. So, without any adjustments, you blindly take the last period's value and then assume that to be the forecast for the current value. By the way, such forecasts are used only for comparison with the forecasts generated by better or more sophisticated techniques. So, let us say if you are forecasting a particular ARMA model using the minimum MSC criteria, then, towards the end of the day, one can actually compare the forecast that you get from the minimum MSC criteria with a naive forecast.

So, essentially, a naive forecast is nothing but guessing, right? Because if you already have some value, you are basically guessing as to what the value would be in this month, okay? Alright. Now, in this period, we have a U statistic given by Theil's. So, it is called the Theil's U statistic.

Now, again, do not go into too much detail, but then here you see two different quantities. So, U1 and then U2, and then expressions for that, right? And again, I mean, if you pause the video for a second or something like that, then the expressions are not very difficult. So, what exactly is U1? So, U1 is nothing but the square root of squared errors, right?

Because AT is the actual value, FT is the forecasted value. You are taking the square error, then summation, and then under root. And then in the denominator, you have the individual squared value. So, summation a t square plus summation f t square, both under root. And then u2 is slightly different.

So, u2 is one step ahead in the future, if you observe, right? So, what is u2? So, u2 is nothing but you are involving f t plus 1, a t plus 1 also divided by 80, and then the denominator is kind of only based on the actual values, ok. So, for any given data or for any given practical set of observations, one can actually find out the numerical value of U1, right, and the numerical value of U2, obviously, once you implement some forecasting technique or a forecasting measure, right. Now, these are all the properties of the Theil's U statistics.

$$U_1 = \frac{\sqrt{\sum_{t=1}^{n}(a_t - f_t)^2}}{\sqrt{\sum_{t=1}^{n} a_t{}^2} + \sqrt{\sum_{t=1}^{n} f_t{}^2}}, \quad U_2 = \sqrt{\frac{\sum_{t=1}^{n-1}\left(\frac{f_{t+1} - a_{t+1}}{a_t}\right)^2}{\sum_{t=1}^{n-1}\left(\frac{a_{t+1} - a_t}{a_t}\right)^2}}$$

So, if U1 is close to 0 or if U1 is closer to 0, it indicates better forecasting accuracy. So, if my U1 is really really close to 0, it indicates better forecasting accuracy. On the other hand, if U2 equals 1, there is no difference between a naive forecast and the forecasting technique that you have applied. So, the idea of the Theil's U statistic is to kind of compare the forecast given by the technique you have applied to a naive forecast. And then these are some fixed measures for that.

So, whenever either U1 or U2 takes up some value, then what would happen to the forecasting technique and the naive forecast, right? For example, if U2 is 1, there is no difference between the naive forecast and the technique. So, the forecasting technique reduces to guessing. On the other hand, if U2 is less than 1, the technique is better than a naive forecast. And lastly, if U2 is greater than 1, the technique is worse than a naive forecast.

So, I think Theil's U statistic could be very much applied to compare the forecast given by the technique you have applied. Let us say minimum MSE or whatever technique you apply as compared to guessing or as compared to a naive forecast. So, probably in the next session this week, we will explore more ideas about, let us say, forecasting in general. So, let us say some smoothing techniques and then Holt's method, Holt-Winters' approach, and then towards the end, we will take up some practical data and try to connect all the ideas we have studied this week and then try to implement that on the data. Thank you.